IÉSEG SCHOOL OF MANAGEMENT

MARCH 2022

Introduction to Statistical and Machine Learning

A report on Statistical and Machine Models
Individual Project

PRESENTED TO
Minh Phan

PRESENTED BY
Enita Omuvwie

Table of Contents

INTRODUCTION

This project was based on data from a bank marketing campaign. The primary objective is to predict whether the clients will subscribe. I will be using five machine learning methods and it will have a comprehensive understanding as to how the models perform and their advantages and drawbacks. This report entails the basic knowledge of unsupervised learning for pattern recognition and supervised learning for prediction.

Machine learning is a branch of Artificial Intelligence which uses techniques in statistics to provide computers with the ability to learn without being programmed which could be progressive to improve the overall performance of the model with preprocessed data that is fed into the system. With this data, the computer will train the data using different models or algorithms to predict a certain outcome or output value. The data fed in would be Input + Output, which is then run on the machine during training, and the machine creates its own program (logic), which could be evaluated during testing, this is the logic that differentiates machine learning from traditional programming. Machine learning is classified widely under four categories which are Supervised learning, Unsupervised learning, Semi-Supervised learning and Reinforcement learning only three will be explained below.

SUPERVISED LEARNING

The goal of supervised learning is to receive inputs and outputs are provided to create a match between the input and output variables as the desired value. The training procedure is repeated until the model reaches the required level of accuracy on the training data. The output is either categorical which is classification method or regression for numerical output values. Examples are image classification, market regression.

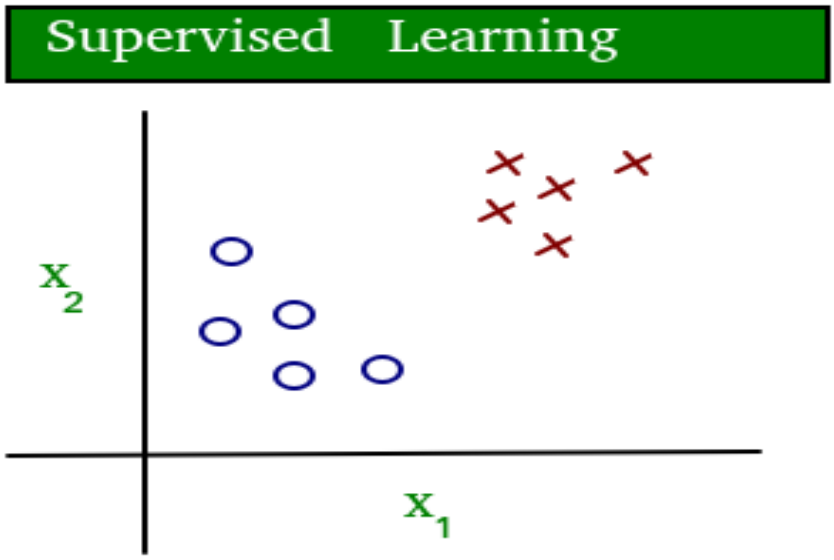

Figure 1:This shows supervised learning using classification

UNSUPERVISED LEARNING

This modeling method restructures the data into something unique, including such new features which may indicate a class or a new series of statistically independent values. It is not given any sort of desired output label, instead the learning algorithm focuses on learning through the structure of the unlabeled input data. It can help humans understand the meaning of data and provide new useful inputs to supervised learning methods. It is widely used for segmenting customers into groups for a given strategy by clustering the population into different groups and for building recommendation systems.

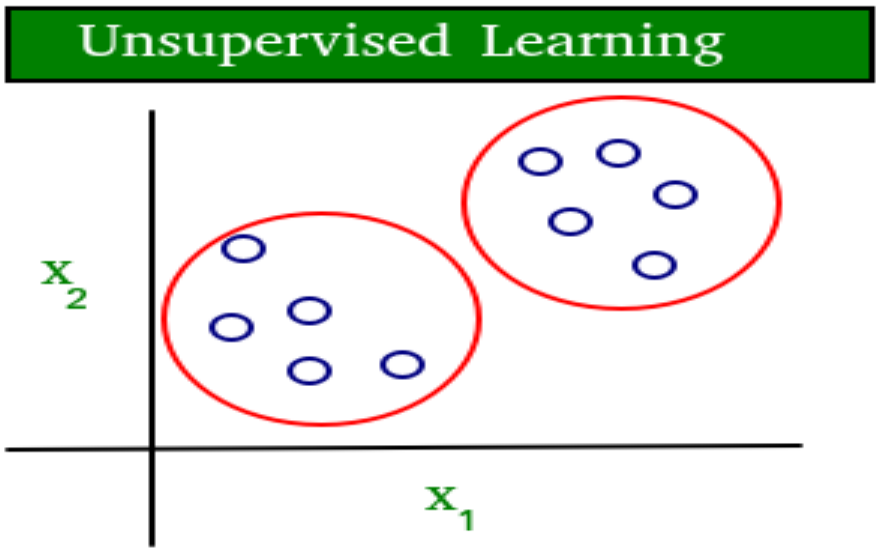

Figure 2:This shows unsupervised learning using clustering

REINFORCEMENT LEARNING

The machine is trained to make specific decisions in this algorithm. It is primarily exposed to an environment in which it continuously learns and trains through trial and error. It attempts to learn from experience or knowledge to improve the model and make better decisions based on these findings. Unlike unsupervised learning, the data fed into reinforcement learning models usually does not have labels, and as a result, the model provides either positive or negative feedback that can be descriptive or prescriptive in nature. An example will be Google's DeepMind reinforcement (Atari) which helps to play old video games by itself.

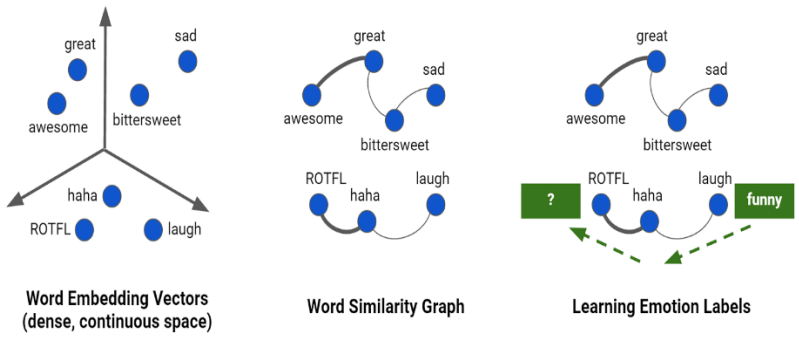

Figure 3:This shows reinforcement learning using word vectors

There are tasks which are performed based on the model's purpose and they include:

- Classification: The input variables are divided into one or more classes where the model needs to assign unseen inputs as multiclassification labels for the classes. In this case, classifying customers into two groups of subscribers and nonsubscribers where are inputs are given information about clients in a supervised manner.
- Regression: It is used for estimating real values based on continuous variables, this is a supervised problem which is carried out with historical data.
- Association Rules learning (or dependency modelling): This helps to find the relationship between the input variables. This can be used in market research or market basket analysis to study customer shopping habits.
- Clustering: This divides a set of input variables into groups, and this is an unsupervised learning task. This is used in science research work.
- Dimensionality Reduction: This helps a computer to visualize high dimension of data. Inputs are simplified by mapping them into a lower-dimensional space. An example will be topic modelling for discovering similarity in documents by matching similar topics.

DATA DESCRIPTION

The dataset is a bank marketing csv file containing 20 independent variables and 1 dependent variable. It has a total of 20000 observations with a total of 10 categorical variables and 10 numeric variables which are a mixture of socio-economic attributes, campaign information and demographics of clients. Below is the data dictionary for the dataset.

S/N	FEATURE	DESCRIPTION	DATATYPE	REMARK
1	client_id	This contains the client unique identifier	int64	
2	age	It shows the age of the client	float64	
3	job	This contains the job type	object	Possible values include job titles: admin., housemaid, management etc.
4	marital	It holds the marital status	object	Values like divorced, married, single
5	education	Shows the education level of customers	object	Values like basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree
6	default	This shows the status of credit default for clients	object	Either yes or no
7	housing	This shows the status of housing loan for clients	object	Either yes or no
8	loan	This shows the status of personal loan for clients	object	Either yes or no

9	contact	It contains contact communication type for campaign	object	Either cellular or telephone
10	month	It contains last contact month of year	object	From jan to dec
11	day_of_week	It contains days of the week for last contact of customers	float64	From mon to sun
12	campaign	It has the number of contacts performed during the campaign for clients	float64	
13	pdays	Difference between last contact date and a previous campaign	float64	999 signifies was not contacted previously
14	previous	Contains the number of contacts performed before this campaign and for the customer	float64	
15	poutcome	It has previous marketing campaign outcome	object	Possible values are failure, nonexistent, success
16	emp.var.rate	It is employment variation rate — quarterly indicator	float64	
17	cons.price.idx	It is consumer price index — monthly indicator	float64	
18	cons.conf.idx	It is consumer confidence index — monthly indicator	float64	
19	euribor3m	It is euro inter-bank offer 3-month rate — daily indicator	float64	
20	nr.employed	It is number of employees — quarterly indicator	float64	
21	subscribe	It contains the status of subscription to a term deposit	int64	Either 1 or 0

EXPLORATORY DATA ANALYSIS STEPS

1. Reading in the csv file containing the dataset
2. Data cleaning steps were applied which included checking for missing values and outliers
3. Carry out exploratory data analysis steps by checking the summary, size, info of the data
4. Then fill missing values with mean of each of the columns
5. Dummy and label encoding processes were carried out to create new features and change categorical variable
6. Created a new basetable with the 35 new variables and 20000 observations
7. Created the correlation matrix to check the relationship between variables

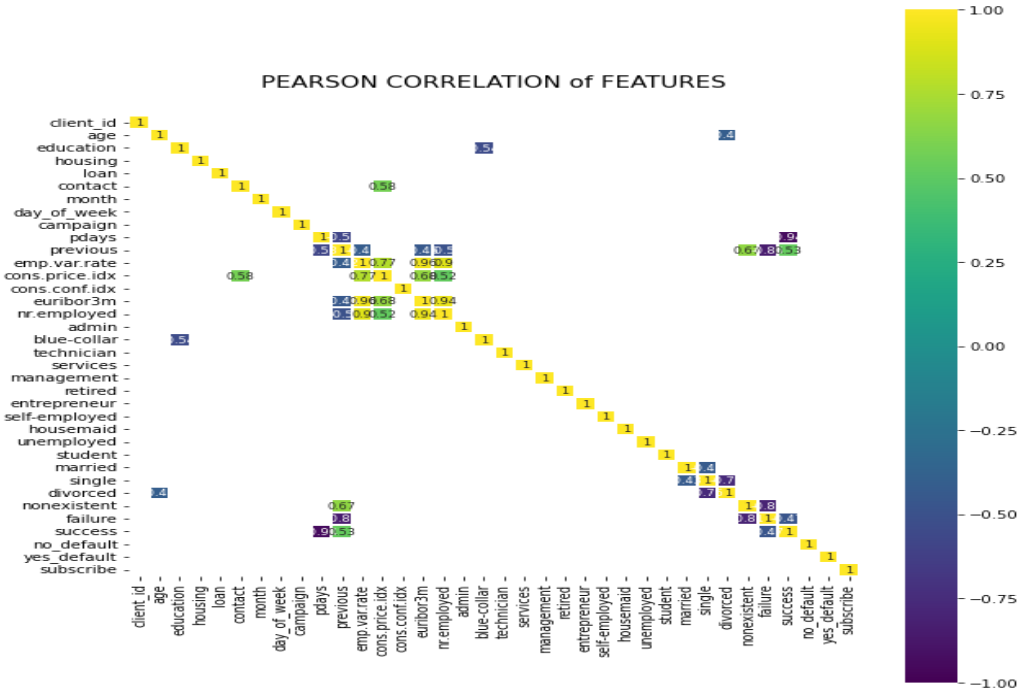

Figure 4: This shows the correlation matrix between variables

MACHINE LEARNING ALGORITHM

The objective of this approach is to predict the Subscription of bank clients with given variables in the data set. When the correlation has been carried out then splitting can be done into training and test dataset for model evaluation. There are two methods that can be used for the above step:

- Regression: this is used to predict the numerical value of subscription. The metrics are Mean Absolute Error, which is the average difference between predictions and true values, and R2, which is the percentage of variation in the response explained by the model.
- Classification: This includes classifying the data into groups and predicting the class. Accuracy, f1 score (the harmonic mean of precision and recall), and the confusion matrix are the metrics.

LINEAR REGRESSION

This is used to predict the quantitative response of Y from predictor variables X. It assumes there is a linear relationship between X and Y with this it performs a logistic task. This regression model checks the relationship between the dependent and independent variables putting into consideration the number of independent variables that is fed into the system, with this it models a target prediction value. The expression for the relationship is

$$Y \approx \beta 0 + \beta 1 X.$$

When training a model, we are provided with the variables X and Y. These symbols are interpreted as:

X: input variable for training data

Y: labels to the data

β0: represents the intercept

β1: represents the slope

When training the model, we use the values of X to fit the best line to predict the value of Y. The model uses the best values of β0 and β1 to get the best-fit regression line. The model aims to predict y value in such a way that the error difference between predicted value and true value is as small as possible by achieving the best-fit regression line. As a result, it is critical to update the β0 and β1 values to achieve the best value which minimizes the error between the predicted y value (pred) and the true y value (true) (y). The RSS(Residual Sum Square) can be defined as the amount of variability that is left behind after regression. It can be minimized using the Least Square Coefficient Estimates. Residual Mean Square Error (RMSE) is the difference between predicted y value (pred) and true y value (y). RMSE helps to provide an absolute measure of lack of the fit of the model to the data.

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

… Equation for RSS

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

… Equation for LCSE

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}.$$

… Equation for RMSE

There are other ways to measure fit of a model, these are briefly explained below:

The R^2 statistic is a different way of measuring fit. It is expressed as a proportion—the proportion of variance explained—and thus always has a value between 0 and 1, and is independent of the scale of Y. If close to 1 it denotes large proportion of variability and 0 indicates that the model is wrong. It has more advantage in terms of interpretability over RMSE.

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

… Equation for R^2

Another method is the use of F-statistics to test for the relationship between the response and predictors where p-value is relatively minute compared to a value sample size n for variable selection. Three selection methods are forward, backward, and mixed selection.

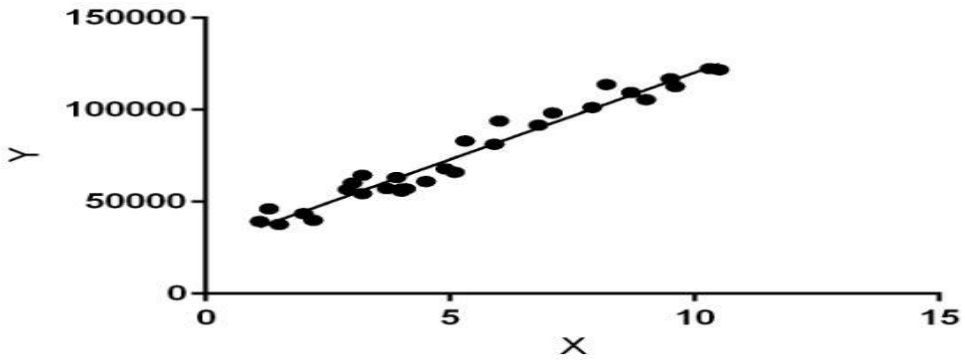

Figure 5: Simple Logistic Regression using line of best fit

ADVANTAGES

- It is easy to implement and interpret the output variable coefficient.
- It is less complex when the algorithm knows the relationship between independent variables and dependent variables.
- Since the algorithm is likely to experience overfitting, there are methods that can be used to avoid it e.g., cross validation, dimensionality reduction and regularization.

DISADVANTAGES

- It looks at the mean of the independent variables and dependent variables which is not a complete description of a single variable.
- It assumes that there is a straight-line relationship between the independent and dependent variables.
- In this technique, the boundaries are linear, and the outliers influence the regression. Some other problems are collinearity, high-leverage points, non-linearity of the response-predictor relationship and non-constant covariant.

SUPPORT VECTOR MACHINE

SVM is a parametric classifier characterized explicitly by a separating hyperplane. It is a supervised learning model with associated learning algorithms for data classification and regression analysis. Given labeled training data (supervised learning), the algorithm generates an optimal hyperplane for categorizing new examples. This is represented as of the examples as points in space, mapped so that the examples of the different categories are separated by as wide a gap as possible. It uses the kernel to enlarge feature space in a specific manner to accommodate a non-linear boundary between classes. The kernel is an efficient computational approach in such situation.

In this algorithm, each data item is plotted as a point in n-dimensional space (for which n is the number of features), with the value of each feature being the value of a specific coordinate. In a n-dimensional space, a hyperplane is a flat affine subspace of dimension n – 1. A hyperplane is defined by the equation for two-dimension:

$$\beta 0 + \beta 1X1 + \beta 2X2 = 0 \ldots \text{Equation for simple hyperplane}$$

for parameters β0, β1, and β2. For any X = (X1,X2)T for which the equation holds are a point on the hyperplane. It simply is the equation of a line, which in two dimensions a hyperplane is a line. Our aim is to create a classifier based on training data that correctly classifies test observations based on feature measurements. If we classify our test observation x based on the sign of $f(x^*) = \beta 0 + \beta 1x^*_1 + \beta 2x^*_2 + \ldots + \beta px^*_p$. The observation is assigned to class 1, if f(x*) is positive, and if f(x*) is negative, it is assigned to class -1. If f(x*) is far from zero, then x is far from the hyperplane, and we can be sure about the class assignment for x. The magnitude of f(x*) can also be used.

We use the maximal margin hyperplane or optimal separating hyperplane is a separating hyperplane that is far from the training observation, and it has the largest minimum distance. The margin is the smallest distance between the observation and the hyperplane and with this we can calculate the distance from each of the observations and a separating hyperplane. The classification of the test observation based on the side

of the maximal marginal hyperplane it falls is what we consider as maximal margin classifier. The rule applies as follows that if a classifier has a large margin on the train data it will also have it on the test data. The maximal margin hyperplane is directly dependent on the support vectors and is solves the problem of optimization. When construction the margin, we set constraints to make sure each observation is on the correct side of the hyperplane and least distance margin from the hyperplane.

Another method is the support vector classifier which is more robust to individual observations and better classifies the training observations. It is classifying the test observation based on the side of the hyperplane it falls and the hyperplane is used to separate the observations correctly into classes which may also be misclassified by a few data points. It uses a tuning parameter to manage the tolerance of the hyperplane and margin to know the number of violations it can handle.

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

… Equation for Linear SVM kernel

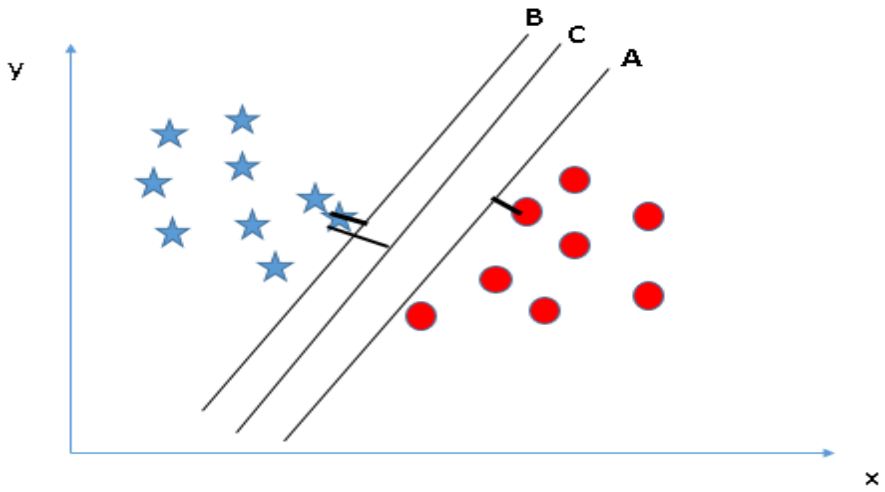

Figure 6: The middle line is the margin, both lines A and B are hyperplanes to classify both data points

It quantifies the similarity of observations using Pearson correlation to compare them in pairs. This helps to create a fit is an improvement from the support vector classifier. Sometimes, the feature space makes computations unsurmountable, imperative, and vastly complex.

ADVANTAGES

- It works well in three-dimensional spaces.
- It is also memory efficient because it uses a subset of training points in the decision function (called support vectors).
- It works extremely well when there is a clear margin of separation.
- It works well when the number of dimensions is greater than the number of samples.

- It works on non-linear problems, and it is not biased on outliers.

DISADVANTAGES

- Feature scaling is an important procedure.
- It is not easy to understand so it is not commonly used.
- When we have a large amount of data, it tends not to perform well because the required training time is longer.
- It also does not perform well when the data set contains more noise, i.e., target classes overlap.
- SVM does not directly provide probability estimates; these are obtained through an expensive five-fold cross-validation procedure.

GRADIENT BOOSTING CLASSIFIER MODEL

Boosting is a machine learning algorithm that aids in the reduction of variance and bias in a classification algorithm ensemble. Ensemble learning entails constructing a strong model from a collection (or "ensemble") of "weaker" models. When compared to a single model, this approach produces better predictive performance.

Gradient Boosting Model (GBM) is a machine learning technique that is used for classification and regression to give a prediction through ensemble of weak prediction models e.g., Decision Tree. The basic idea is to train a set of classifiers (experts) and then let them vote. When dealing with a large amount of data, GBM is used to make a prediction with high predictive power. It does not modify the sample distribution.

The model has three primary elements:

- Additive Model: This uses a step-by-step incremental approach to add trees(weak learning), one each period, as it is done iteratively, we get closer to the final model. With each iteration, there is a reduction in the value of the loss function.
- Loss Function: This is used to guesstimate how accurate a model is at making predictions based on the data. It defers depending on the nature of the problem.
- Weak Learner: This is the classifying principle for the data which does poorly. It is seen as a high error rate, and it is also referred to as decision stumps.

The initial step is using the constant value where:

$$F_0(x) = argmin \sum_{i=1}^{n} L(y_i, \gamma)$$

… Equation for constant values

y_i : observed values

L: loss function

Gamma: value for log(odds)

This adds up all the loss functions for the observed values then find a log value that will minimize the sum and calculate the residual value. Then we can fit the regression tree and create terminal regions. For each leaf on the new tree, the gamma is calculated as output. The below equation is used approximate the loss function. The third equation is used to calculate the output of the tree.

$$L(y_1, F_{m-1}(x_1) + \gamma) \approx L(y_1, F_{m-1}(x_1)) + \frac{d}{dF()}(y_1, F_{m-1}(x_1))\gamma + \frac{1}{2}\frac{d^2}{dF()}(y_1, F_{m-1}(x1))\gamma^2$$

… Equation for Taylor Polynomial

$$\gamma = \frac{Residual}{p * (1-p)}$$

… Equation for gamma

$$\gamma = \frac{Sum\ of\ residuals}{Sum\ of\ each\ p(1-p)\ for\ each\ sample\ in\ the\ leaf}$$

… Equation for GBM

$$Update\ F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

…Equation for updating predictions

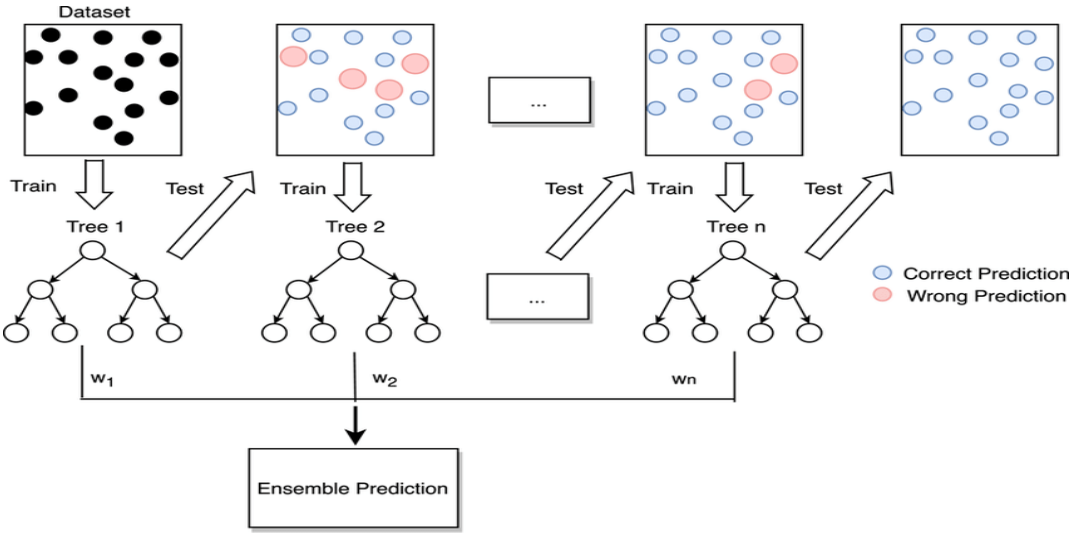

Figure 7:This shows prediction from a dataset using the gradient boosting

ADVANTAGES

- It handles data preprocessing steps and missing data.
- It works well for categorical and numerical data.
- It provides prediction accuracy that cannot be trumped.

- It is very flexible as it can optimize loss function and can provide some hyperparameter tuning options.

DISADVANTAGES

- It is not interpretable, but tools can be used for this.
- It requires grid search tuning due to its high flexibility nature which may cause heavy impact on the model.
- It has high computational power which makes it expensive as it consumes memory and time.
- It may sometimes overfit and overemphasize outliers due to its ability to minimize errors. Some methods to help this are tree constraints, randomized sampling, shrinkage, and penalized learning.

RANDOM FOREST MODEL

It is a supervised learning method that builds its trees from an ensemble of decision trees. It is used to classify objects based on a random selection of attributes at each node for it to split when this is being carried out each tree votes and the most popular class is returned. This method starts by creating multiple subsets from the original data set. Each subset of features is used to create an iteration of node split. As the tree grows, the whole processing is repeated taking into consideration the aggregated value of predictions for n value of predictions. For each tree planted, the N data point is taken at random from the training set with replacement at different points. The number of trees p is decided, the tree is constructed, and predictions are made based on the mean value of p on the test set.

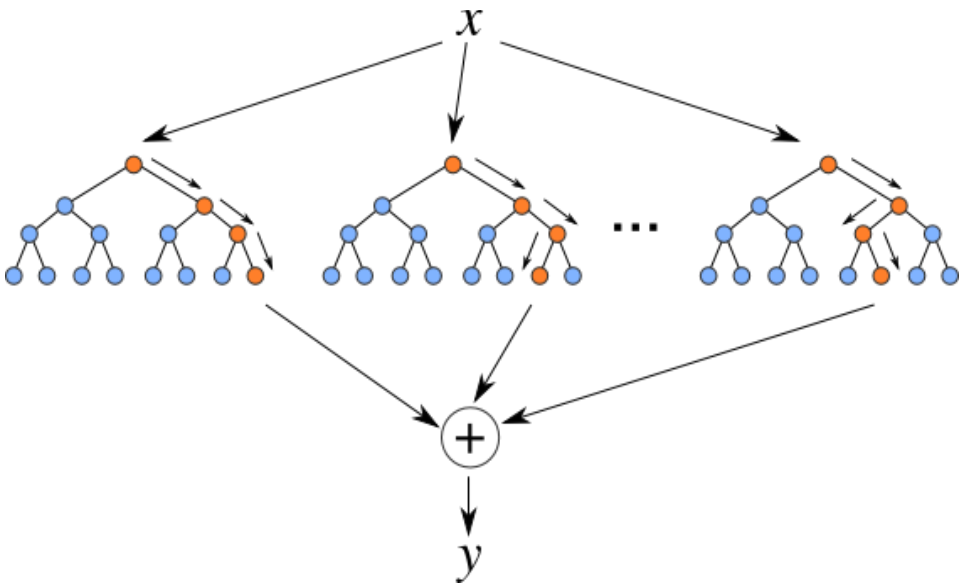

Figure 8:Figure shows tree construction and selection to predict a y value on Random Forest

ADVANTAGES

- It can be used for both regression and classification task and works well on both numerical and categorical data.
- It is used to perform feature selection and create uncorrelated decision trees as it chooses random set of features to build each tree.

- It is not influenced by outliers as it does binning on the variables.
- It can work on linear and nonlinear relationships well and provide high accuracy without bias.

DISADVANTAGES

- It is not interpretable as it is a Blackbox model and gives less control to the user.
- It is very expensive as it uses high computation for its activities.
- Overfitting can occur easily, and the number of trees is to be determined by the user.

K NEAREST NEIGHBOUR ALGORITHM

K Nearest Neighbour is a supervised learning model that does both classification and regression. KNN is used to predict the class in test observation to classify the nearest observations. It uses labelled data points to predict a new data point class by voting on its k neighbours for classification problems while in regression problems, it uses the median or average of the continuous values in KNN to predict the continuous value for the new class of data point. Nearest neighbours are the smallest distance points between our new class of data and other data points in the feature space while K basically is the number of points considered when we apply our algorithm. This is measured by a distance function which can either be Euclidean, Manhattan, Hamming and Minkowski. Hamming is used mainly in cases that have only categorical variables while the others are continuous functions. It is useful for resampling and applying the correction to missing values. In as much as this is used before some models like SVM and ANN as a standard, it is usually advisable to scale variables and carry out preprocessing steps to avoid bias, noise, and outliers.

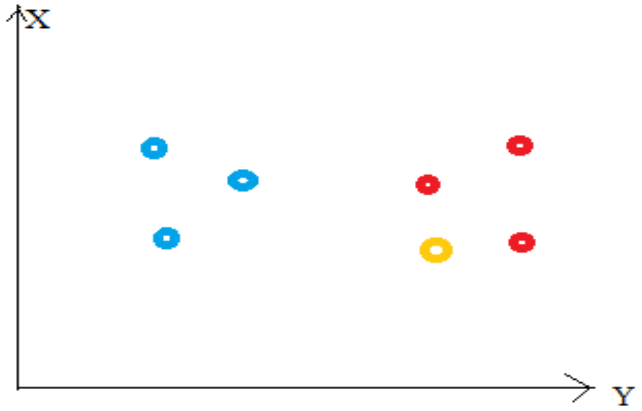

Figure 9: This shows a typical example of classification in KNN, if both red and blue are classes of friends and yellow is a new school kid, for yellow to mix up it will check the nearest class which in this case is red

This model uses an error curve to measure the level of high variance which is also called overfitting. This can be seen in the graph below, that from the instance the error in the test data due to high variance and stabilizes onwards due to the increase in K value. Other points to note in the model include domain knowledge for the choice of K values and in the case of binary classification, the K value should be odd. For the preprocessing steps to avoid mentioned errors, standardization or normalization can be performed (data scaling steps), feature selection, PCA (principal component analysis) both are common dimensionality

reduction steps and attending to the missing values either by replacing them with mean/mode or deleting rows involved.

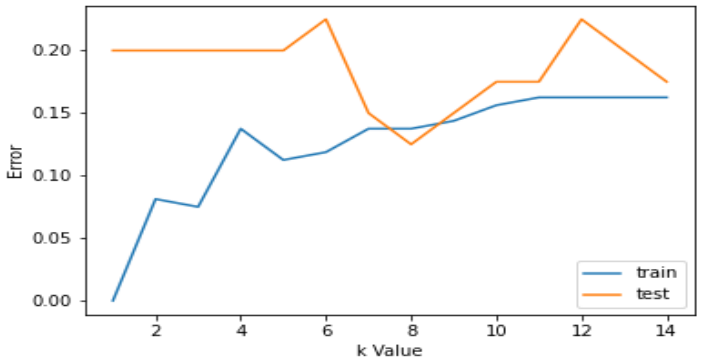

Figure 10: This shows the error curve for the training and test dataset

ADVANTAGES

- It is simple and convenient to use as it makes use of the K nearest neighbours for the classification of new data points in a data set even in multiclass datasets.
- This algorithm is not parametric like regression it does not assume what should be done prior to implementation like training and building of a model.
- It is a memory-based approach and responds if the algorithm makes changes as it is used in real-time.
- It can be used for both classification and regression problems and it uses hyperparameter tuning.

DISADVANTAGES

- It fails and grows slow as the dataset size increases and it applies as the number of input variables increases, it has difficulties predicting new output variables in the dataset.
- It is very sensitive to outliers and has no ability to handle missing variables.
- It has a problem if the data is imbalanced and if the optimal number of neighbours is not set, it has a problem with data entry.
- It always requires the same scaling for the features while developing the model, so all attributes need to be treated the same way.
- It has high computational power which makes it expensive.

REFERENCES

Advantages and Disadvantages of different Regression models - GeeksforGeeks. (2022). https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-regression-models/?ref=lbp

Boosting. (2022). https://corporatefinanceinstitute.com/resources/knowledge/other/boosting/

Classifying data using Support Vector Machines(SVMs) in Python - GeeksforGeeks. (2022). https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-python/

EDA/eda.ipynb at main · broepke/EDA. (2022). https://github.com/broepke/EDA/blob/main/eda.ipynb

Getting started with Machine Learning - GeeksforGeeks. (2022). https://www.geeksforgeeks.org/getting-started-machine-learning/

Gradient Boosting for Classification | Paperspace Blog. (2022). https://blog.paperspace.com/gradient-boosting-for-classification/

IESEG\Communication Tools\Comm Tools Group Assignment.ipynb. (2022).

IESEG\Stat Machine Learning\Example_v5.3 - Group 1 - Data Processing\MBD2021_InClass Kaggle_2_Modeling_Group1_Py_v5.3 - Data Processing.ipynb. (2022).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning. Springer.

KNN Algorithm | What is KNN Algorithm | How does KNN Function. (2022). https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/

ML | Linear Regression - GeeksforGeeks. (2022). https://www.geeksforgeeks.org/ml-linear-regression/

Pros and Cons of K-Nearest Neighbors - From The GENESIS. (2022). https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/

Random Forest: Pros and Cons. (2022). https://medium.datadriveninvestor.com/random-forest-pros-and-cons-c1c42fb64f04

Roepke, B. (2021, December 4). My Goto process for exploratory data analysis with python. Data Knows All. https://www.dataknowsall.com/eda.html

Roy, S. (2020, January 18). Machine learning case study: A data-driven approach to predict the success of bank telemarketing. Medium. Retrieved March 31, 2022, from https://towardsdatascience.com/machine-learning-case-study-a-data-driven-approach-to-predict-the-success-of-bank-telemarketing-20e37d46c31c

SVM | Support Vector Machine Algorithm in Machine Learning. (2022). https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Top 4 advantages and disadvantages of Support Vector Machine or SVM. (2022). https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107

WillKoehrsen. (2018, May 21). Machine-learning-project-walkthrough/machine learning project part 1.ipynb at master·Willkoehrsen/machine-learning-project-walkthrough. GitHub. https://github.com/WillKoehrsen/machine-learning-project-walkthrough/blob/master/Machine%20Learning%20Project%20Part%201.ipynb