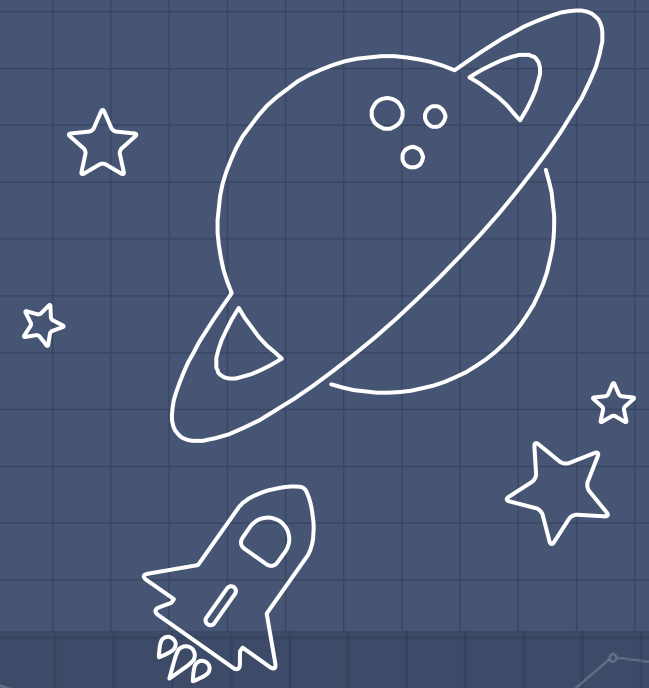
arise:

2019 Data Science Internship

Powered by:

paylater



SUPERVISED LEARNING: CLASSIFICATION

Powered by:

LEARNING OUTCOMES

- Differentiate between supervised and unsupervised learning
- Understand the process of building a binary classification model
- Hands-on practice in Binary Classification

TERMINOLOGIES

1. Dataset: A collection of related set of records (rows) for which certain variables (columns) have been measured and observed.
2. Labelled Dataset: A labelled dataset is a set of observations for which each observation has been tagged as belonging to a group from a given set of possible groups.
3. Unlabelled Dataset: A labelled dataset is a set of observations for which each observation has not been tagged as belonging to any group.

Dividing data into 'training' and 'test' sets is an important part of model developing. This is done mostly for evaluation purposes.

1. Training Data: Training data (usually 60-80% of the data) is the part of the data used for developing models. The models are built using relationships observed between variables only in the training data.
2. Testing Data: The test set (usually 20-40% of the data) is the part of data used for testing the performance of the model. The idea is to test if models developed with training data can pick up the same relationships in new data it has not seen before.
3. Discrete Variable: A variable that can take a finite (only one of a known set of values) number of values. E.g Gender (Male or Female), States (any one of 37 states) etc.
4. Continuous Variable: A continuous variable can take an infinitely number of values e.g weather readings, stock prices, weight, height.

WHAT IS SUPERVISED LEARNING?

Supervised learning is a method used to enable machines to classify objects, problems or situations based on the learn relationships between features of historical data fed into the machines.

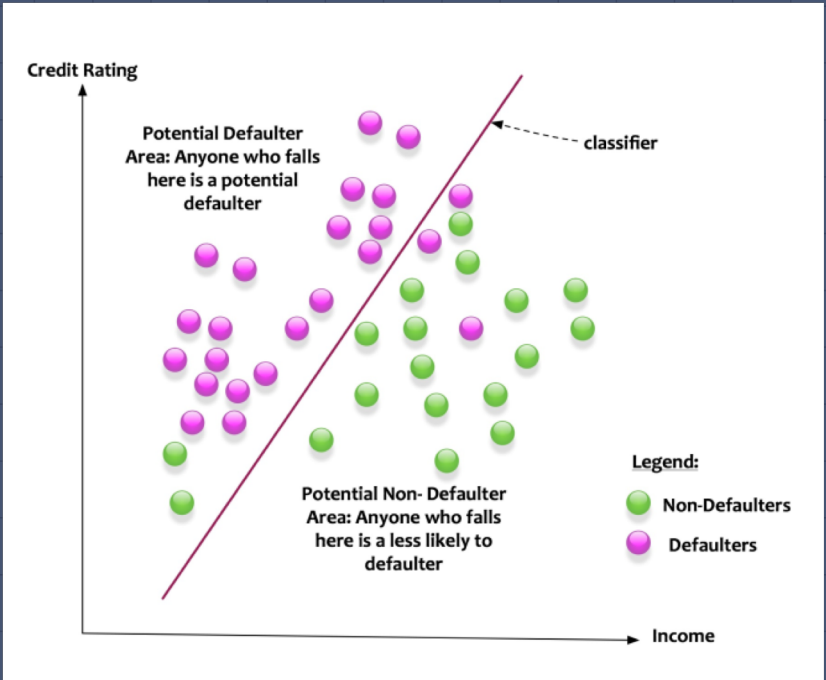
The idea behind supervised learning can be thought of as ‘Learning From Data’.

THE GRAND QUESTION?

Given the observed relationships between attributes of an object or event that qualifies them to belong to a certain group, can we develop a representative model that allows us to infer what group a new object or event will fall into given the same features?

Supervised learning is often done in form of **‘classification’** when we want to map the relationship between a set of input variables to a discrete output variable e.g Died/Survived OR **‘regression’** when we want to map the relationship between a set of input variables to a continuous output variable e.g weather readings.

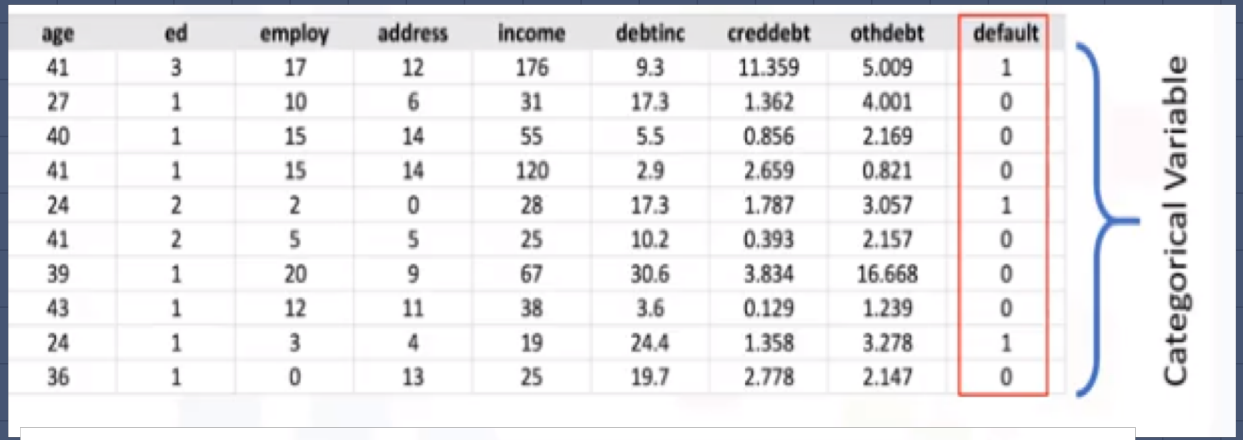
CLASSIFICATION



- A supervised learning approach.
- Categorizes some unknown items into discrete set of categories or classes.
- The number of possible classes in which items can fall into determines the ‘n-class’ classification problem; where n >= 2.
- A data set with n=2 possible classes of objects (see image beside) is called a ‘Binary Classification’ .
- A data set with n>2 possible classes objects is called a ‘multi-class’ classification.

HOW CLASSIFICATION WORKS

Classification predicts the class label for an unlabelled test case



age	ed	employ	address	income	debtinc	creddebt	othdebt	default
41	3	17	12	176	9.3	11.359	5.009	1
27	1	10	6	31	17.3	1.362	4.001	0
40	1	15	14	55	5.5	0.856	2.169	0
41	1	15	14	120	2.9	2.659	0.821	0
24	2	2	0	28	17.3	1.787	3.057	1
41	2	5	5	25	10.2	0.393	2.157	0
39	1	20	9	67	30.6	3.834	16.668	0
43	1	12	11	38	3.6	0.129	1.239	0
24	1	3	4	19	24.4	1.358	3.278	1
36	1	0	13	25	19.7	2.778	2.147	0

Categorical Variable

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
37	2	16	10	130	9.3	10.23	3.21	

READING

https://medium.com/seismic-data-science/how-classification-works-51d61c675b6e

HOW CLASSIFICATION WORKS II

Binary (also called **binomial) classification** is the task of classifying the elements of a given set into two groups (predicting which group each one belongs to) on the basis of a **classification rule**.

A **classification rule** is a defined criteria of which observations in a dataset must meet before they can belong to a 'group'.

The grouping variable of which observations can belong to is called a **'target'** variable. In this example, the target variable is 'defaulted' because that is what we are trying to predict.

AGE	INCOME (000$)
41	176
27	31
45	120
24	28
38	55

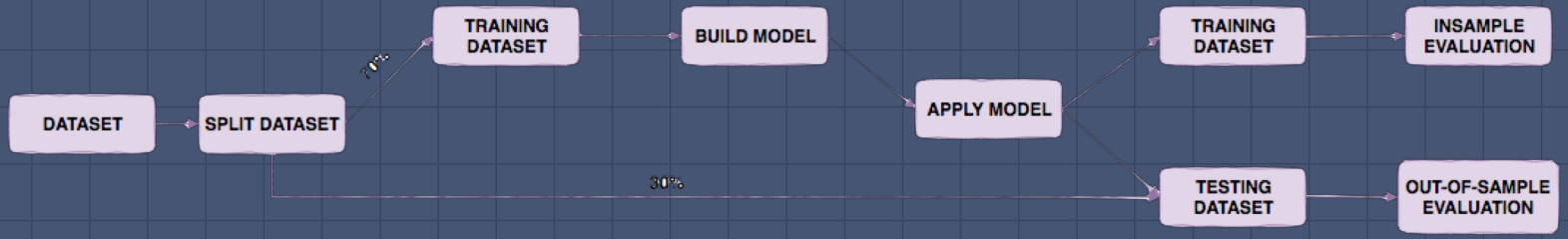
AGE	INCOME (000$)	DEFAULTED
41	176	1
27	31	0
45	120	1
24	28	0
38	55	1

Based on this data, let's build a simple classifier. We decide that

Classifier: If age < 50 and income is < $40,000 customers will default on their loans but if age >=50 and income >= $50,000, customers will repay their loans. The resulting predicted values are

All classification problems simply require you to build the best classifier that correctly predicts the target variable (defaulted).

THE CLASSIFICATION MODEL BUILDING AND EVALUATION PROCESS



1. All classification problems begin with a labelled dataset.
2. The dataset is then split into 2 (training and testing sets)
3. The model is developed using the training data
4. The model is used to predict the target variable in the training set
5. The model is also used to predict the target variable in the test set.
6. To evaluate the accuracy of the model, we compare the predicted values of the train (in sample accuracy) and the test set (out of sample accuracy)

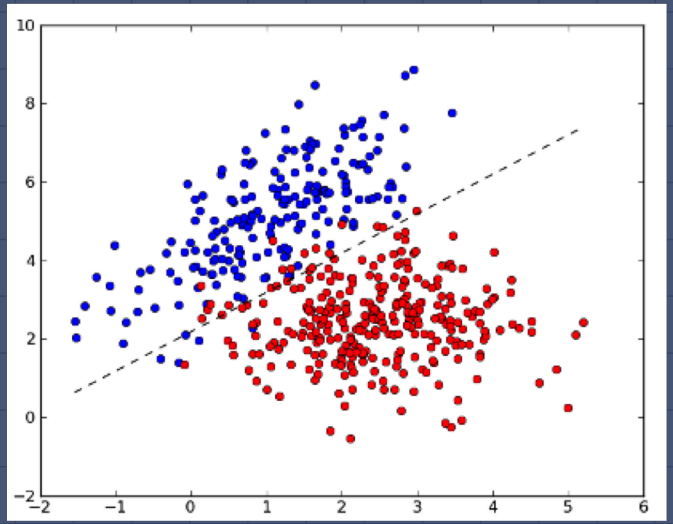
HANDS-ON (CLASSIFICATION): Who Survived the Titanic?

https://www.kaggle.com/sashr07/kaggle-titanic-tutorial

BINARY (BINOMIAL) CLASSIFICATION

Binary or binomial classification is the task of classifying the elements of a given set into two groups on the basis of a classification rule.

The goal is to develop a classifier that draws a line through the data separating it into two groups.



EXAMPLES OF BINARY CLASSIFICATION PROBLEMS

1. Who will default on their loan payment (Paid/Default)?
2. Who will survive the titanic (Died/Survived)?
3. Who will change their telecommunications provider (Churn/Not-Churn)?
4. Who will graduate or drop out of university

LOGISTIC REGRESSION (BINARY CLASSIFICATION)

Regression analysis is a set of statistical processes for estimating the relationships among variables.

Although there are various methods for binary classification, we will cover basic Logistic Regression in this course.

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis.

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables (As seen in the examples above).

Logistic Regression : Building a logistic regression model step by step

HANDS-ON: VARIOUS TYPES OF CLASSIFICATION

1. Decision Trees: Decision trees in python
2. Support Vector Machines : svm in python
3. Random Forests : Random forests in python
4. Neural Networks : Neural networks in python

HANDS-ON: IMPLEMENTING AND EVALUATING THE VARIOUS BINARY CLASSIFICATION METHODS

https://www.kaggle.com/klaudiajankowska/binary-classification-methods-comparison