LET'S LOOK AT THE PROBLEM OF

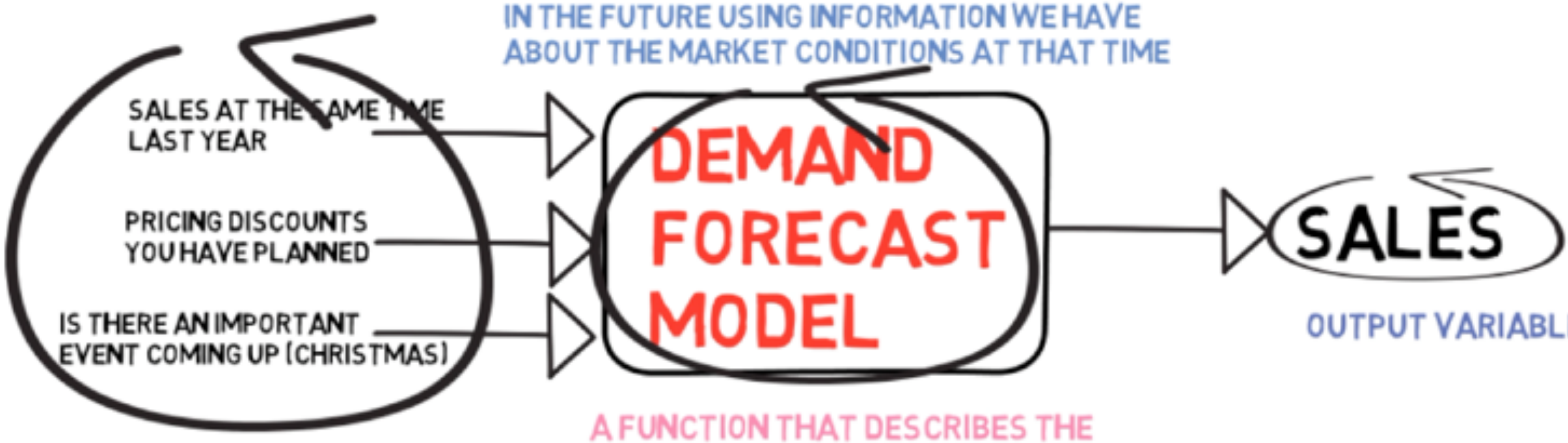
DEMAND FORECASTING

IN MANY BUSINESSES IT IS CRITICAL TO ESTIMATE WHAT THE SALES AT A GIVEN POINT IN THE FUTURE MIGHT BE

IN MANUFACTURING, THIS WOULD BE USED TO PLAN AHEAD FOR THE PRODUCTION CYCLES

A RETAILER WOULD USE IT TO BUILD UP/SCALE DOWN THE INVENTORY THEY NEED TO STORE

A TYPICAL SOLUTION FOR DEMAND FORECASTING WOULD INVOLVE PREDICTING SALES AT A POINT IN THE FUTURE USING INFORMATION WE HAVE ABOUT THE MARKET CONDITIONS AT THAT TIME



SALES AT THE SAME TIME LAST YEAR

PRICING DISCOUNTS YOU HAVE PLANNED

IS THERE AN IMPORTANT EVENT COMING UP (CHRISTMAS)

DEMAND FORECAST MODEL

SALES

INPUT VARIABLES

OUTPUT VARIABLE

A FUNCTION THAT DESCRIBES THE RELATIONSHIP BETWEEN THE INPUT VARIABLES AND OUTPUT VARIABLE

DEMAND FORECASTING IS A PROTOTYPICAL APPLICATION OF A CLASS OF MACHINE LEARNING ALGORITHMS CALLED

REGRESSION

REGRESSION

REGRESSION IS CONCERNED WITH MODELLING THE RELATIONSHIP BETWEEN VARIABLES

THE INPUT VARIABLES ARE USUALLY CALLED INDEPENDENT VARIABLES OR PREDICTORS

$(X1, X2, X3, \ldots, Xn)$

THE OUTPUT VARIABLE IS CALLED THE DEPENDENT VARIABLE Y

THE REGRESSION ALGORITHM WILL TRY TO FIND A FUNCTION THAT CAN COMPUTE THE PREDICTED VALUE OF Y GIVEN THE INPUTS

$$Yp = F(X1, X2, X3, \ldots, Xn)$$

THIS IS DONE BY LOOKING AT A TRAINING DATA SET WHICH HAS SOME KNOWN INPUT, OUTPUT PAIRS (USUALLY TIME SERIES DATA OF PAST EVENTS)

THE REGRESSION ALGORITHM THEN TRIES TO **MINIMIZE THE ERROR** FOR THE TRAINING DATA

THE DIFFERENCE BETWEEN THE ACTUAL VALUE Y AND THE PREDICTED VALUE Yp

IT IS OFTEN IMPORTANT TO UNDERSTAND THE PROBABILITY DISTRIBUTION OF THE ERROR. SPECIFIC REGRESSION TECHNIQUES MAKE ASSUMPTIONS ABOUT WHAT THE DISTRIBUTION IS, AND ONLY WORK IF THOSE ASSUMPTIONS AR

SINCE REGRESSION INVOLVES AN EXPLICIT TRAINING STAGE **IT IS A FORM OF SUPERVISED LEARNING**

LINEAR REGRESSION

LINEAR REGRESSION ASSUMES A LINEAR RELATIONSHIP BETWEEN THE DEPENDENT VARIABLE AND INDEPENDENT VARIABLES

WHEN THERE IS ONLY 1 INDEPENDENT VARIABLE IT IS KNOWN AS SIMPLE LINEAR REGRESSION

REGRESSION CO-EFFICIENT

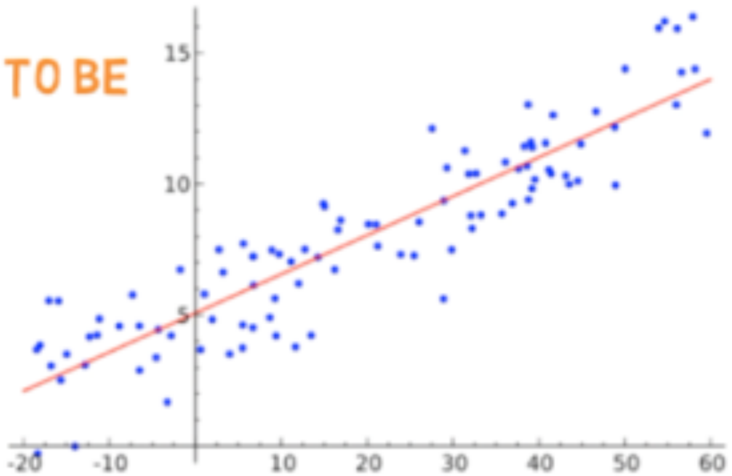
$$Y = \beta_0 + \boxed{\beta_1} X + \varepsilon$$

ONE POPULAR TECHNIQUE FOR DOING THIS IS THE ORDINARY LEAST SQUARES METHOD

LINEAR REGRESSION WILL TRY TO FIND A LINE THAT BEST FITS THE Y VALUES

LINEAR REGRESSION IS ADVISED TO BE USED ONLY IF THE ERROR (Y-YP) IS NORMALLY DISTRIBUTED

LINEAR REGRESSION CAN ALSO BE USED WITH MULTIPLE INDEPENDENT VARIABLES - THIS IS - **MULTILINEAR REGRESSION**



LOGISTIC REGRESSION

THIS IS USED WHEN THE DEPENDENT VARIABLE IS CATEGORICAL IE. IT CAN ONLY BE ONE OF A FIXED SET OF VALUES (RED,BLUE,GREEN)

GIVEN THE DEPENDENT VARIABLES, LOGISTIC REGRESSION PREDICTS THE PROBABILITY OF EACH OUTCOME.

THE INDEPENDENT VARIABLES (PREDICTORS) CAN BE CONTINUOUS OR CATEGORICAL

LOGISTIC REGRESSION IS STILL A LINEAR CLASSIFIER, BECAUSE THE BOUNDARY THAT IT DRAWS IS BASED ON FINDING A LINEAR FUNCTION OF THE INDEPENDENT V

$$P = \frac{e^{\boxed{a+bX}}}{1+e^{a+bX}}$$

FOR EXAMPLE, THE PREDICTORS COULD BE AGE AND GENDER, THE OUTCOMES COULD BE "ADMITTED TO COLLEGE" / "NOT ADMITTED"

LOGISTICS REGRESSION OFTEN WORKS WELL AS A CLASSIFICATION APPROACH (ASSIGN PROBLEM INSTANCE TO THE OUTCOME WITH THE HIGHEST PROBABILITY)