PREDICTIVE MODELLING

PREDICTIVE MODELLING

▪Predictive modeling is the problem of developing a model using historical data to make a prediction on new data where we do not have the answer.

▪Predictive modeling can be described as the mathematical problem of approximating a mapping function (f) from input variables (X) to output variables (y).

▪This is called the problem of function approximation.

▪The job of the modeling algorithm is to find the best mapping function we can given the time and resources available.

CLASSIFICATION
PREDICTIVE MODELLING

CLASSIFICATION PREDICTIVE MODELLING

▪Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

▪The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation.

▪For example, an email of text can be classified as belonging to one of two classes: "spam" and "*not spam*".

CLASSIFICATION PREDICTIVE MODELLING

▪A classification problem requires that examples be classified into one of two or more classes.

▪A classification can have real-valued or discrete input variables.

▪A problem with two classes is often called a two-class or binary classification problem.

▪A problem with more than two classes is often called a multi-class classification problem.
▪A problem where an example is assigned multiple classes is called a multi-label classification problem.

▪An algorithm that is capable of learning a classification predictive model is called a classification algorithm.

CLASSIFICATION PREDICTIVE MODELLING

▪It is common for classification models to predict a continuous value as the probability of a given example belonging to each output class. The probabilities can be interpreted as the likelihood or confidence of a given example belonging to each class. A predicted probability can be converted into a class value by selecting the class label that has the highest probability.

▪For example, a specific email of text may be assigned the probabilities of 0.1 as being "spam" and 0.9 as being "not spam". We can convert these probabilities to a class label by selecting the "not spam" label as it has the highest predicted likelihood.

▪There are many ways to estimate the skill of a classification predictive model, but perhaps the most common is to calculate the classification accuracy.

REGRESSION
PREDICTIVE MODELLING

REGRESSION PREDICTIVE MODELLING

▪Regression predictive modeling is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y).

▪Regression is different from classification, which involves predicting a category or class label.

REGRESSION PREDICTIVE MODELLING

▪Regression refers to predictive modeling problems that involve predicting a numeric value.

▪It is different from classification that involves predicting a class label. Unlike classification, you cannot use classification accuracy to evaluate the predictions made by a regression model.

▪Instead, you must use error metrics specifically designed for evaluating predictions made on regression problems.

REGRESSION PREDICTIVE MODELLING

▪A continuous output variable is a real-value, such as an integer or floating point value. These are often quantities, such as amounts and sizes.

▪For example, a house may be predicted to sell for a specific dollar value, perhaps in the range of 100,000to200,000.

▪A regression problem requires the prediction of a quantity.

▪A regression can have real-valued or discrete input variables.

▪A problem with multiple input variables is often called a multivariate regression problem.

▪A regression problem where input variables are ordered by time is called a time series forecasting problem.

REGRESSION PREDICTIVE MODELLING

▪Regression predictive modeling are those problems that involve predicting a numeric value.

▪Metrics for regression involve calculating an error score to summarize the predictive skill of a model.

▪How to calculate and report mean squared error, root mean squared error, and mean absolute error.