Machine Learning Concepts

# Data analytics

Data Analytics refers to the techniques to analyze data to enhance productivity and business gain.

It is a process of cleaning and transforming and modeling data with the goal of discovering some useful information and support decision making.

Data analyst will collect data from various sources, analyze the data and generate reports to provide hidden insights to business.

1. **Descriptive analytics**

* help you understand what happened
* insight into the past
* Use Descriptive Analytics when you need to understand at an aggregate level what is going on in your company, and when you want to summarize and describe different aspects of your business.

1. **Diagnostic analytics**

* why something happened
* “How can we avoid this problem” or “How can we duplicate this solution?” These are covered by diagnostic analytics.

1. **Predictive analytics**

* what could happen - understanding the future
* Use Predictive Analytics any time you need to know something about the future, or fill in the information that you do not have.

1. **Prescriptive analytics**

* what should we do
* advise on possible outcomes
* Use Prescriptive Analytics anytime you need to provide users with advice on what action to take.

# Predictive Analysis:

* A branch of advanced data analytics which is used to make predictions about unknown future events.
* Uses data, statistical algorithms and Ml techniques to identify the likelihood of future outcomes based on historical data.
* These models use known results to develop (or train) a model that can be used to predict values for different or new data.
* The patterns found in historical data can be used to identify risks and opportunities for future.
* Predict trends, understand customers, improve busieness performances
* EX- commonwealth bank can predict likelihood of the fraud activity for any given transaction before it is authorized with 40 milli seconds.
* Allows org to become proactive, forward looking, anticipating outcomes and behaviors based on data and not on assumptions.

## Predictive analytics process

1. Define project – identify the data sets, defining the project outcomes
2. Data collection – data from multiple sources for analysis
3. Data analysis: cleaning, transforming and modelling data for discovering useful information.
4. Statistics – validating assumptions, hypotheses using standard statistic models.
5. Modelling - creating the best model
6. Deployment – deploy the analytical results in to the everyday a decision making process
7. Monitoring – Reporting/ Web reporting

# Machine Learning Techniques:

Machine learning is a subset of AI which provided machines the ability to learn automatically & improve from experience without being explicitly programmed.

Machine Learning classified in to

1. Supervised learning (is a method in which we teach machine using labelled data)
2. Regression
3. Classification
4. Un-supervised learning (machine is trained using unlabeled data and without any guidance)
5. Reinforcement Learning (in this learning, an agent interacts with its environment by producing actions & discovers errors or rewards)

# Regression Analysis

**Regression**: it is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. By performing regression analysis we can confidently tell which factor matter most, which factor can be ignored, and how these factors influence each other.

* Form of a predictive modelling technique which tells the relationship b/w a dependent (target) and independent variables (predictors).
* Tells 2 S, S- significant Relationship b/w dependent and independent, another S indicates Strength of impact on dependent by independent variables.
* This method is used for forecasting, time series modelling and finding causal effect relationship b/w the variables.
* Ex – relationship b/w rash driving and number of accidents by a driver is best studied through regression.

**Types of Regression:** are based on 3 N’s

* Number of independent variables
* Nature of the dependent variable
* Nature of the regression line

**Linear Regression Logistic Regression**

**Polynomial Regression Stepwise Regression**

**Ridge Regression Lasso Regression, Elastic Net Regression**

# Variables in Statistics

**Quantitative vs Categorical Variables**

Quantitative **–** numerical variables

Ex: Numbers, count, percent

Categorical – Any variable that is **not quantitative** is categorical.

Ordinal (ordered list) (very low, low, medium, great, very great),

Nominal (unordered list) (male/female, married/unmarried/divorcee)

Ex: Gender, Yes or no, hair color,

**Discrete vs Continuous Variables (both comes in Quantitative Variables)**

If a variable can take on **any value between two specified values**, then it is called a continuous variable; otherwise, it is called a discrete variable.

Continuous - Variable whose value is obtained by **Measuring**

Ex: Time, height, weight, Age, income

Discrete – variable whose value is obtained by **Counting**.

Ex: Number of plates, number of heads,

**Note**: Interval variable is as same as Ordinal but the intervals b/w the variables are equally spaced. Interval and ratio are comes under Numerical variables opposite to categorical one.

# Linear Regression:

* Dependent variable is continuous.
* Independent variable (s) can be continuous or discrete
* Nature of the regression line is linear
* Solves Regression problems.
* Linear relationship b/w target and predictor is must.

Linear Regression establishes a relationship between dependent variable (Y/target) and Independent variables (X/Predictor) using a best fit straight line (regression line).

Represented by

**Y=mx+C**

Where C= Intercept, m is the slope of the line

This equation can be used to predict the value of the target variable using predictor variables.

**Multiple linear regression** - has >1 predictor variables

## Line of best fit:

The vertical distance between the actual data points to the best fit line are called errors.

Main idea is to fit this line of regression by minimizing the sum of squares of these errors.

This is also known as **Principle of least square method**

Best Regression line will be obtained by using **Least Square method**

**C= (Y\_mean) – m (X\_mean)**

**m= Sum ((X-X\_mean) (Y-Y\_mean)/ (X-X\_mean) (X-X\_mean))**

**Definition:** It calculates the best fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to line.

All regression lines should pass through means of X, Y.

## Accuracy of the model:

### R2 method:

Coefficient of Determination

Tells how well our predicted values are matching with the actual values.

Mathematically it is a ratio of predicted values and observed values.

**Definition**: it measures the proportion of the variation in your dependent variable(y) by all your independent variables (x) in the model. It assumes that every independent variable in the model helps to explain variation in the dependent variable. In other words it assumes that all the independent variables will help to predict our target.

**It lies between 0% and 100%**

**Higher R vaule, better the model**

**R2 = sum ((Y\_pred – Y\_mean) ^2) / sum ((Y\_actual-Y\_mean) ^2)**

If R2=1 means it is a best fit

R-squared is a relative measure of fit

### Adjusted R ^2

This is a form of R-Squared that is adjusted for the number of terms in the model.

**Definition**: it tells us the proportion of the variation explained by only those independent variables that really help in explaining the dependent variable.

R\_adjusted ^2 = 1- ((1-R^2) (n-1/n-p-1))

Where R^2 = R-squared of the model

n = sample size

p = number of predictors in the model

This index is extremely useful to determine possible overfitting in the model.

* Every time you add an independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines. Whereas Adjusted R-squared increases only when independent variable is significant and affects dependent variable.
* Adjusted R-squared value always be less than or equal to r-squared value.
* Adjusted R-Squared can be negative, if R2 is close to 0.
* If adjusted R square increases when a new variable is added to the model, the variable should remain in the model.
* If the adjusted R square decreases when the new variable is added then the variable should not remain in the model.

### Root Mean Square Method:

RMSE tell you how well the actual data points are concentrated around the line of best fit.

RMSE is an absolute measure of fit.

Lower values of RMSE indicate better fit.

MSE = Mean Squared Error = 1/n Sum (Y\_actual – Y\_predicted) ^2

RMSE = Root Mean Square = Sqrt (MSE)

SSE = Sum of Squared error = n\*MSE **SSE = n\*MSE**

**SSE tells how far the data are from the model predicted values**

**Sum of Squares of the differences between the observed values and the expected values**

R^2 = 1-(SSE/TSS)

TSS = Total sum of squares = Sum(Y\_actual – Y\_mean) ^2

# Logistic Regression:

* Dependent variable is Categorical (Binary in nature (0/1, yes/No, True/False)).
* Value of Y ranges from 0 to 1.
* Nature of the regression line is S-Curve (Sigmoid function curve).
* Solves Classification problems
* Does not require linear relationship b/w dependent and independent variable.
* If Dependent variable is Ordinal, Ordinal Logistic Regression.

Y\_predicted = b0+b1x1+b2x2+…bnxn

# Polynomial Regressoin:

* Power of independent variable is >1
* Y=a+bx^2
* Best fit is a curve.

Stepwise Regression:

* We will use when multiple independent variables are involved.
* Selection of the independent variables are done by an automatic process.
* Stepwise Regression methods

1. Standard Stepwise Regression – adds and removes predictors as needed for each step.
2. Forward Selection – Starts with the most significant predictor in the model and adds variable each step.
3. Backward Elimination – starts will all predictors in the model and removes the least significant variable for each step.

Ridge Regression:

* When data suffers from multicollinearity (predictors are highly correlated)