The Present & Future of Data Analytics

Dr. Siddhartha Roy

Goals of this presentation

- Provide Analytics overview
- Discuss various types of Analytical models
- Highlight current and future trends
- Analytics project Lifecycle and Workflow
- Analytics Landscape

Table of Contents

What is Analytics?

Analytics is defined as a scientific process which helps organizations in:

Answering business questions and taking fact-based decisions

Identifying patterns

Discover hidden relationships

Forecasting and Predictions

Understand market trends

Define customer preferences

What do Experts say…

“ Without data, you are just another person with an opinion”

– W. Edward Deming eminent researcher & scholar

“Data are becoming the new raw material of business.”

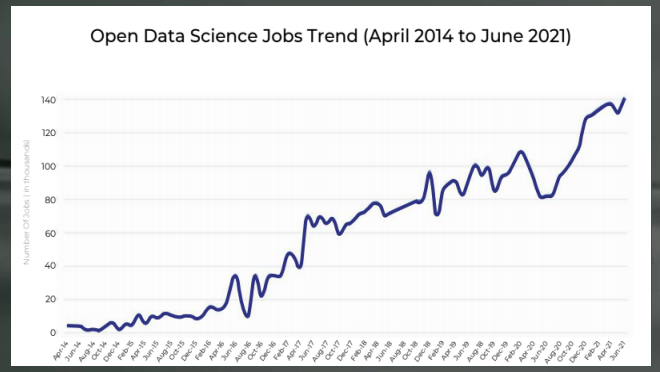
– Craig Mundie Senior Advisor to the CEO at Microsoft

“We're making this analogy that AI is the new electricity. Electricity transformed industries: agriculture, transportation, communication, manufacturing.”

– Andrew Ng Co-founder and Head of Google Brain

Facts about Data Analytics Jobs in India

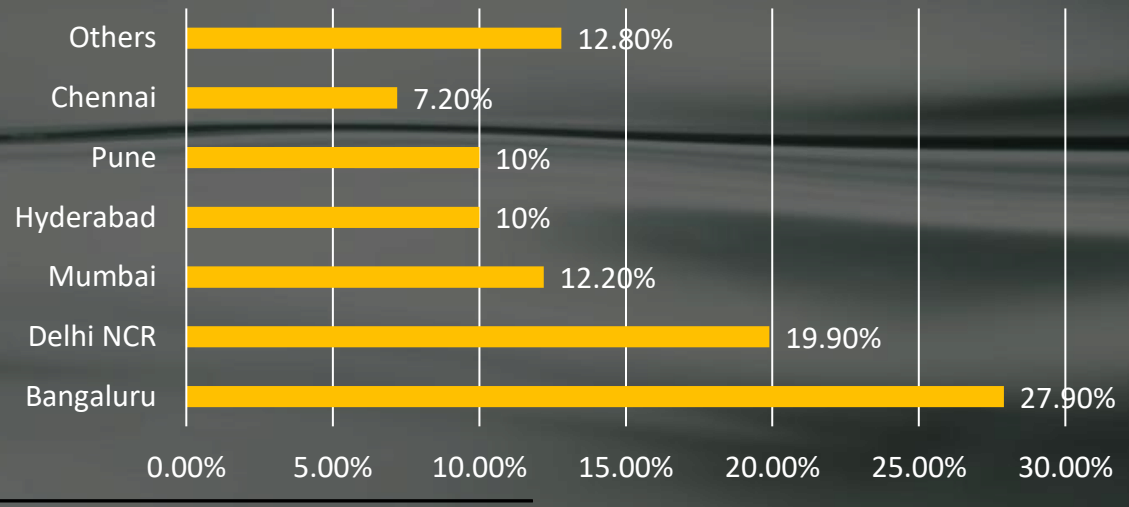
Job Opportunities



47.1% increase in open jobs as compared to last year. Around 137K open positions as per Jun-21

India contributed to **9.4%** of the total global analytics job openings

Share of Open Jobs by City



Jobs by Sector/Industry

Top Sectors with share of Open Jobs June 2021

1. BFSI **27.8%**
2. Energy & Utilities **17.3%**
3. Media & Entertainment **17.3%**
4. Pharma & Healthcare **16.7%**
5. Ecommerce **11.6%**

Jobs by Tools & Technologies

- Python, R, and Tableau are key skill-sets required in Analytics jobs
- AWS, Azure, & Google Cloud are offering cloud platform jobs
- Sharp increase in demand for Spark & Hadoop skills

Source: Analytics India Magazine June-2021

Where can we apply Analytics?

- Pharma & Health care
- Retail
- Ecommerce
- Logistics
- Agriculture
- Education
- Defense
- Telecom
- Automobile
- Energy & Utilities
- Travel & Hospitality
- Media & Entertainment

and many other..

Types of data Analytics

Descriptive

- Trends analysis
- Summarization
- Exploratory data analysis
- Historical summary

Diagnostic

- Root cause analysis, why it happened
- how many times it happened
- Evaluating different scenarios

Predictive

- Predict outcomes
- Forecasting

Prescriptive

- Finding optimal solution
- Identify possible actions

Cognitive

Advanced form of analytic

- Artificial Intelligence
- Machine Learning

Analytical Models

Analytical Models

Default rate model

What are the influential factors which are associated with default rates?

Event Modeling

Can I predict consistent models that accurately reflect the business world

Collection Analysis

Can I segment my delinquent portfolio basis DPD, Amt overdue, area of needs & past transactions?

Optimization

How to arrive at the optimum premium amount of products to meet profit targets?

Risk Based Pricing

What is optimum premium or price to be charged to customers basis their profiles & transaction patterns?

Fraud Analytics

How can we detect potential fraudulent transactions and prevent it?

Cross-sell model

How can we cross-sell new products to the existing customers?

Performance Analysis and Scorecard

How to develop performance ranking system for insurance agent?

Analytical Models Continued..

Conversion Analysis

Can I segment customer base basis their conversion rate?

Portfolio Modeling/Analysis

What is the revenue break up by Customer loyalty segment?

Operational Analytics

Can I take out topmost redundant statements in calling script to reduce AHT?

Spend/Usage Analysis

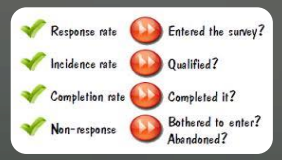
What is the expenditure data with the purpose for reducing procurement costs, improving efficiency and monitoring compliance?

Claim Analytics

How can we analyze claims patterns and process claims efficiently?

ROI analysis

Which are my different project investment and what is my profitability?



Response Rate /Activation Modeling

How to optimize promotional cost by predicting high response segment to any promotion?

Retention /Churn Model

Who are the customers that are going to churn in next cycle & what are the corrective measures to avoid the same?

Analytical Models Continued..

Market Basket Analysis

What are the purchasing patterns of the customers

Customer Satisfaction Analysis

Can I derive key factors influencing CSAT score basis verbatim analysis of chat with my customers?

Survey Analytics

How can I make an effective questionnaire for my survey?

Cross-sell analysis

How can I design a cross-sell strategy for product portfolio?

Product Conversion Rate Analysis

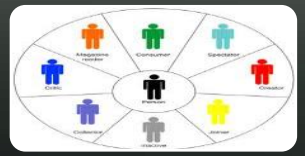
Can I segment my URL basis their conversion rate?

Pricing Optimization

How to arrive at the optimum price of products to meet profit targets?

Customer Life-Time value

What is the transition probability of customers patronizing me basis their transaction pattern?

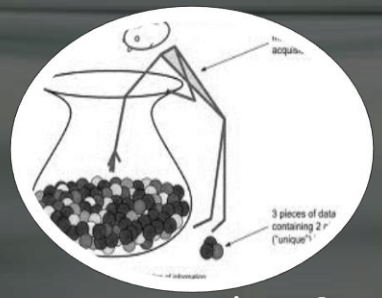


Customer Segmentation

How many discrete customer groups that share similar characteristics basis my profit, demographics etc.

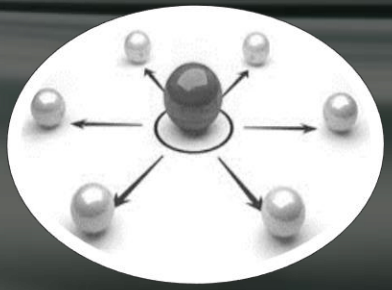
Analytics Workflow

Problem Statement

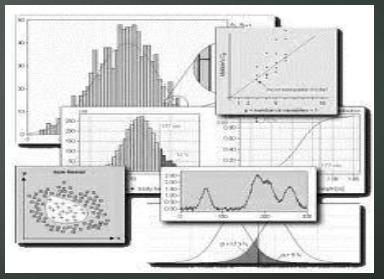


Data Sampling & Consolidation

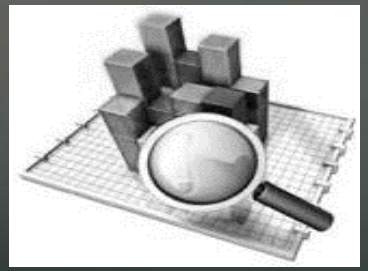
Data Cleaning



Master Data Repository



Exploratory Data Analysis

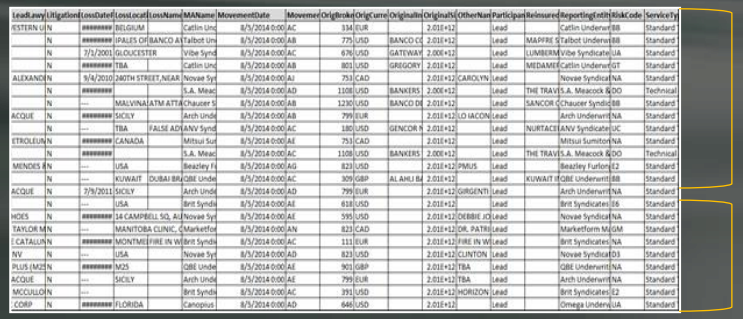


Data Analysis and Modeling

Recommendation

Model Enhancement

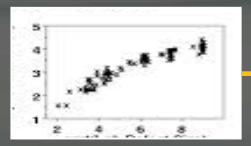
Modeling Journey

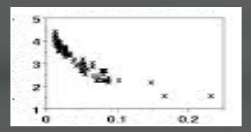


80 %

20 %

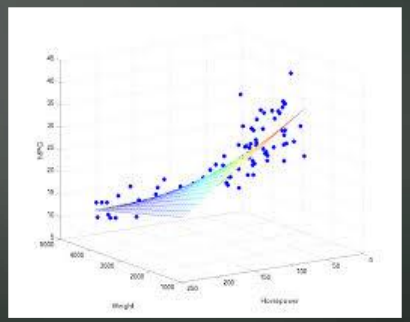
TRANSFORMATION

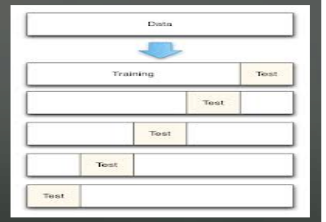




FEATURE SELECTION

MODELING





DIAGNOSTICS CHECK

$$\text{Coefficient of Determination} \rightarrow \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$\text{Sum of Squares Total} \rightarrow \quad SST = \sum (y - \bar{y})^2$$

$$\text{Sum of Squares Regression} \rightarrow \quad SSR = \sum (y' - \bar{y}')^2$$

$$\text{Sum of Squares Error} \rightarrow \quad SSE = \sum (y - y')^2$$

$$Adjusted R^2 = 1 - (1 - R^2)\left(\frac{N-1}{N-k-1}\right)$$

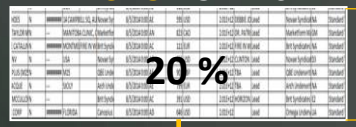
AIC, Cook's distance, Studentized residuals, Standard Residuals

If Not Good

Further refinement and treatment of outliers and variables

If Good

PREDICTION ON TEST DATA



20 %

If accuracy in Train Data ≈ Test Data & Diagnostics are acceptable

PREDICTION ON NEW SET OF DATA

MODEL MONITORING

Analytics Project – Key Phases

1. 
   Opportunity Identification
2. 
   Proof of Concept
3. 
   Delivery
4. 
   Automation, Optimization and Efficiency
5. 
   Large scale Implementation
6. 
   Productization
7. 
   Maintenance & Enhancement
8. 
   Transformation & Innovation

Analytics Landscape

Business Intelligence	Analytics/ Data Science	Big Data Analytics	Artificial Intelligence & Machine Learning
✓ Data cleaning & management ✓ Data reports ✓ Visual Reporting ✓ Creating dashboards	✓ Data Cleaning and management ✓ Exploratory data analysis ✓ Statistical Hypothesis Testing ✓ Predictive Modeling ✓ Forecasting ✓ Identifying patterns ✓ Optimization ✓ Model Diagnostics & Accuracy Testing	✓ Data cleaning & management ✓ Huge amount of data ✓ Social networking sites ✓ Sensors ✓ Structured, semi-structured, unstructured and ✓ Uses parallel processing ✓ Identify fraud analytics ✓ Weather forecasting ✓ Used in designing products ✓ Geographic analysis ✓ Customer Experiences	✓ Data Cleaning & management ✓ Neural networks ✓ Natural Language Processing ✓ Natural language Generation ✓ Image Analysis and Modeling ✓ Sensory Perception ✓ Cognition ✓ Computer Vision ✓ Speech Analytics ✓ Robotics ✓ Expert System ✓ Recommendation Engine ✓ Knowledge Engineering

Job Opportunities

- Data Analyst
- Analytics Associate
- Data Scientist
- Data Engineer
- Researcher
- Business Analyst
- Statistical Analyst/Statistician
- Analytics Consultant
- Analytics Specialist
- Data Analytics Architect
- Data Solution Architect
- Machine Learning Engineer
- Data Analytics Manager
- Marketing Research Analytics
- Operations Analytics Manager

and many more…

Machine Learning

Types of Machine Learning

Supervised Learning

Labeled data is being used for model training both inputs & outputs

- Classification {Categorical Target variable}
- Regression {Continuous Target variable}

Unsupervised Learning

Unlabeled input data, identify structure and patterns from input data

- Clustering {Target variable not available}
- Association {Target variable not available}

Semi-Supervised Learning

Smaller set of labeled data with a large amount of unlabeled data during model training

- Classification {Categorical Target Variable}
- Clustering {Categorical Target Variable}

Reinforcement Learning

Identify an optimal way to accomplish a particular goal

Learns a series of action

- Exploration {Target variable not pre-defined}

Types of Machine Learning

SUPERVISED LEARNING

Classification:

Identifying set of categories sub-populations an observation or observations belongs to

- Logistic Regression
- Decision Trees
- SVM Support-Vector Machines
- Naive Bayes
- K-NN K-Nearest Neighbour

Regression:

Estimate the relationships between a dependent variable often called the 'outcome' or 'response' variable and one or more independent variables

- Linear Regression
- Polynomial Regression
- Ridge/Lasso Regression

SEMI-SUPERVISED LEARNING

Speech Analytics:

Internet Content Classification:

Protein Sequence classification:

[Labelling is time intensive]

UNSUPERVISED LEARNING

Clustering:

Grouping a set of objects in such a way that objects in the same group

- K-means
- Mean-Shift
- Fuzzy C-means

Association Analysis:

Association analysis is the task of finding interesting relationships in large datasets.

- Apriori
- FP-Growth

Dimension Reduction:

Reduction in the number of features or number of observations or both

- Principal Component Analysis PCA
- Partial Least Square Regression PLSR
- Multidimensional Scaling MDS
- Principal Component Regression PCR
- Linear Discriminant Analysis LDA

REINFORCEMENT LEARNING

Q-learning:

Values based learning algo

- Industrial Automation
- Gaming
- Recommendation system
- Chemical reactions
- Self driving cars

R

Programming Language

Introduction to R Programming

- Founded by **Professors Robert Gentleman & Ross Ihaka University of Auckland**
- Open-source programming language
- A whole bouquet of packages
- High quality plots
- Versatile reporting modules
- Free tutorials & learning materials
- Ease of data preparation
- Packages and its applications
  - tidyverse <Data Science>
  - ggplot <Data visualization>
  - dplyr <Data frames>
  - tidyr <Data management>
  - purrr <Functions & Vectors>
  - stringr <String operations>

Packages in R

Library	Description
library(readxl)	**To load data:** These packages help you read and write Microsoft Excel files from R.
library(dplyr)	**To manipulate data:** Provides function subsetting, summarizing, rearranging, and joining together data sets
library(tidyr)	**To manipulate data:** Provide tools for changing the layout of your data sets. Uses the gather and spread functions to convert data into the tidy format
library(gtools)	**To manipulate data:** Provide functions to assist in R programming like calculate the logit and inverse logit transformations, test if a value is missing, empty or contains only NA and NULL values, define macros, sort strings containing both numeric and character components, enumerate permutations and combinations, etc.
library(class)	**Functions for Classification:** Various functions for classification, including k-nearest neighbor, Learning Vector Quantization and Self-Organizing Maps.
library(recipes)	**Preprocessing and Feature Engineering Steps for Modeling:** Prepares data for modeling by providing extensible framework for pipeable sequences of feature engineering steps Statistical parameters can be estimated from an initial data set and then applied to other data sets. The resulting processed output can then be used as inputs for statistical or machine learning models.
library(caret)	**To model data:** Tools for training regression and classification models
library(gmodels)	**To model data:** Programming tools for model fitting.
library(magrittr)	**Data wrangling:** A Forward-Pipe Operator for R that provides a mechanism for chaining commands with a new forward-pipe operator, %>%. This operator will forward a value, or the result of an expression, into the next function call/expression. There is flexible support for the type of right-hand side expressions. This decrease development time and improve readability and maintainability of code.
library(broom)	**Convert Statistical Objects into Tidy Tibbles:** Easy to report results, create plots and works with large numbers of models at once. tidy() summarizes information about model components such as coefficients of a regression. glance() reports information about an entire model, such as goodness of fit measures like AIC and BIC. augment() adds information about individual observations to a dataset, such as fitted values or influence measures
library(plotROC)	**To visualize data:** Generate Useful ROC Curve Charts for Print and Interactive Use
library(ggplot2)	**To visualize data:** Package for making graphics and uses the grammar of graphics to build layered, customizable plots
library(ROCR)	**Visualizing the Performance of Scoring Classifiers:** Tool for creating cutoff-parameterized 2D performance curves by freely combining two from over 25 performance measures

Logistic Regression

Logistic regression

- Supervised algorithm used for classification
- Used for binary classification & can be extended to multi-class classification
- Its equivalent to Linear regression using a sigmoid function
- Models the probability of discreet outcome
- Computes the decision boundaries among the classes
- Used extensively in industry
- Logistic function is represented as

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Background, Data & Objective

Background:

- Heart diseases are leading causes of morbidity and mortality
- Around 17.9 million lives are lost, 85% deaths are due to heart attack & stroke
- Can be attributed to various states – coronary artery diseases, heart rhythm issues, & heart defects
- Cardiovascular disease also refers to narrowed or blocked blood vessels which may result in heart attack, angina or stroke
- A lot of studies related to clinical data analysis involves prediction of cardiovascular diseases
- It is critical to detect cardiovascular disease as early as possible

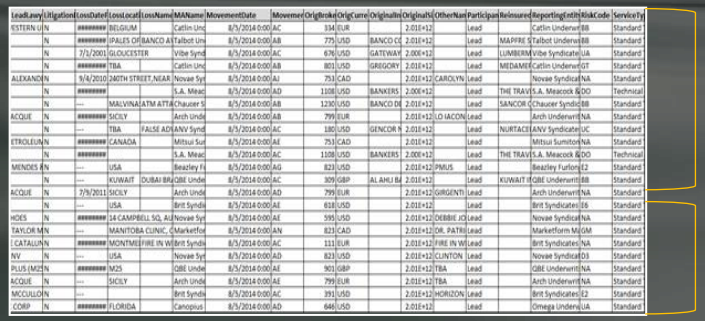
Data:

- The dataset used for this case study is Cleveland heart disease data set
- Original dataset had 76 features but only 14 are chosen here

Objective:

- Build a modeling solution which asses the impact of various factors on heart disease

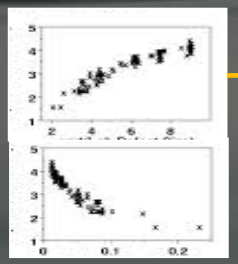
Logistic Regression – Analysis Flow



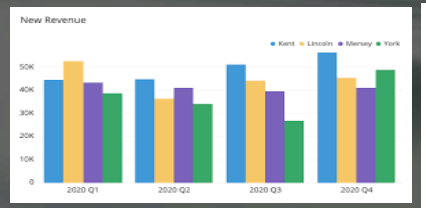
80 %

20 %

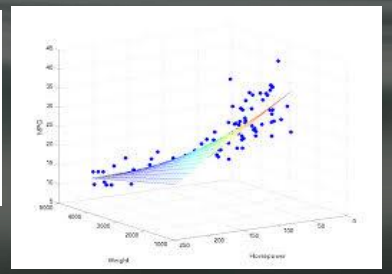
TRANSFORMATION



EXPLORATORY DATA ANALYSIS



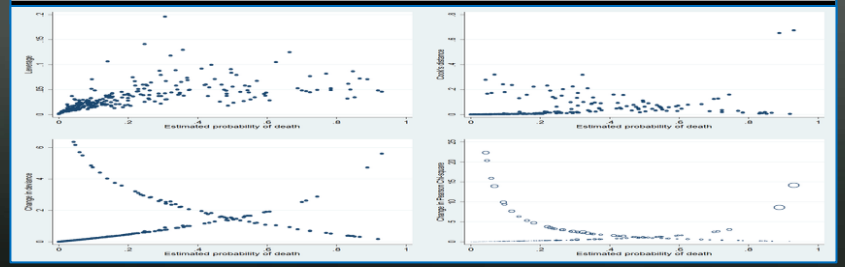
Chi Square, testing of Means, Odds ratio, Scatter plots, etc



VARIABLE SELECTION

DIAGNOSTICS CHECK

	Predicted probabilities	Pearson Chi-square Residuals	Deviance Residuals	Leverages	DFBETA	Delta Deviance	Delta Chi square
Minimum	.07091	-.78954	-.98438	.00672	-.05090	.14772	.0769
Maximum	.38400	3.61982	2.30061	.02656	.32867	5.39994	13.2102
Mean	.17001	.00080	-.17167	.01020	-.00008	.854081	1.0162



Further refinement and treatment of outliers and variables

If Not Good

If Good

PREDICTION ON TEST DATA

Accuracy

Sensitivity, Specificity, ROC curve, Concordance Discordance, Recall Precision

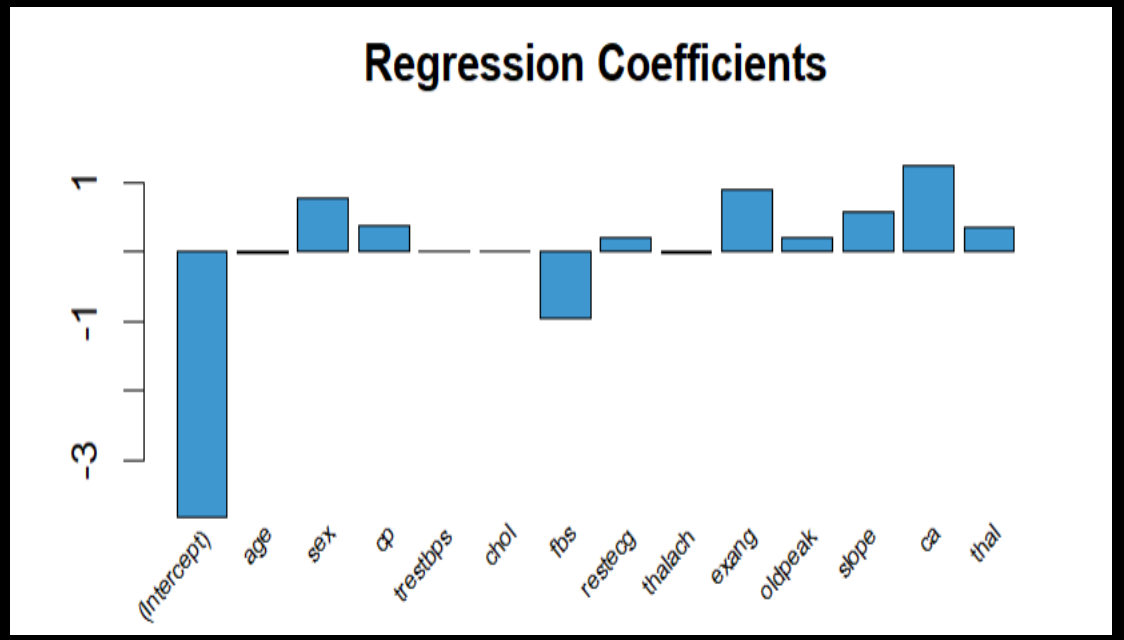
If accuracy in Train Data ≈ Test Data & Diagnostics are acceptable

PREDICTION ON NEW SET OF DATA

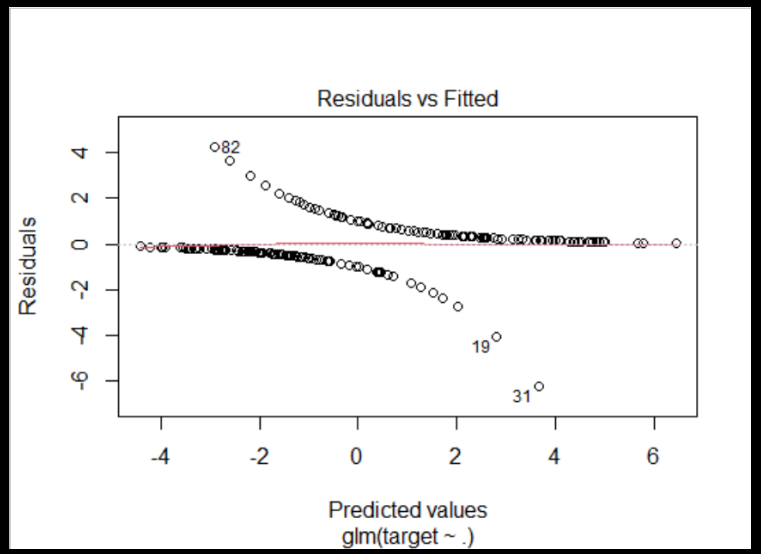
MODEL MONITORING

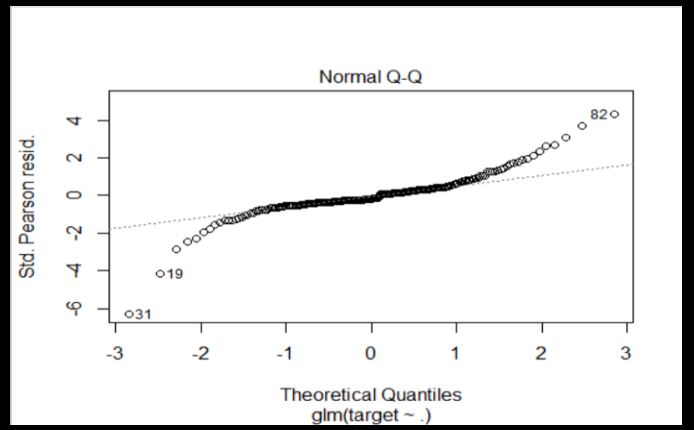
Logistic Regression Results

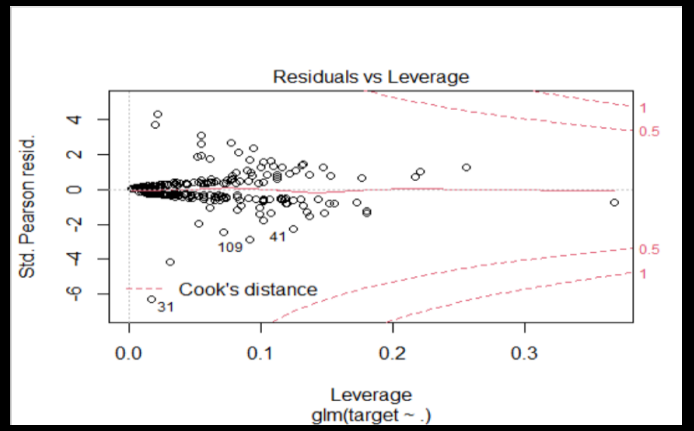
	Estimate	Std..Error	z.value	Pr...z..
(Intercept)	-3.8139086594	3.210250591	-1.1880408	2.348173e-01
age	-0.0271278536	0.028565033	-0.9496875	3.422711e-01
sex	0.7824952931	0.550298666	1.4219466	1.550418e-01
cp	0.3830443844	0.207779892	1.8435104	6.525453e-02
trestbps	0.0162952430	0.012314322	1.3232757	1.857437e-01
chol	0.0006635756	0.004316101	0.1537442	8.778114e-01
fbs	-0.9638660088	0.586327017	-1.6439052	1.001958e-01
restecg	0.2023828626	0.208143136	0.9723254	3.308887e-01
thalach	-0.0198482303	0.011195339	-1.7729013	7.624507e-02
exang	0.8932838663	0.452713203	1.9731783	4.847525e-02
oldpeak	0.2181737222	0.229754532	0.9495949	3.423182e-01
slope	0.5817356136	0.390271310	1.4905928	1.360684e-01
ca	1.2596492928	0.294218497	4.2813396	1.857716e-05
thal	0.3583134114	0.111283940	3.2198124	1.282745e-03

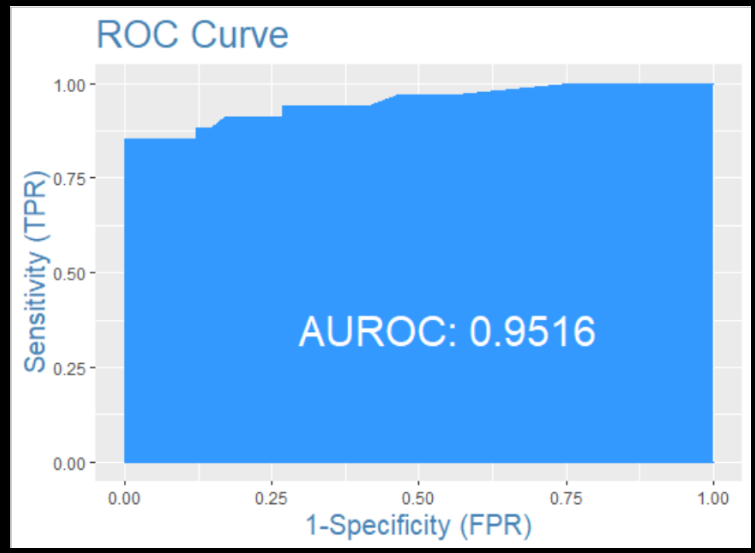


Logistic Regression Results (contd..)









Cluster Analysis

Clustering Analysis

- Unsupervised algorithm
- Used in multiple domains- documents grouping, understanding customer behavior, anomaly detection, etc.
- Segregates the data into subsets with high intra subset similarity and inter-subset dissimilarity
- Few clustering approaches are –
  - Hierarchical clustering,
  - Non- Hierarchy clustering
  - Density based clustering,
  - Partition clustering,
  - Grid based clustering,
  - Correlation clustering, etc.
- K-means is one of the most common algorithm and uses Euclidean distance measures to create optimum clusters

Types of Clustering Techniques

Hierarchical Clustering

{Divides a data set into a sequence of nested partitions – bottom up & top down}

Agglomerative: *Clustering starts with every single object in a single cluster. It merges to the closest pair of clusters as per similarity criteria.*

- The Single-link Method
- The Complete Link Method
- The Group Average Method

Divisive: *Clustering starts with all objects in one cluster and splits large cluster into small pieces*

- DIANA (Divisive Analysis of hierarchical clustering

Partitioning Clustering

{Classifies data into non-overlapping subsets by selecting centroids}

Fuzzy: *Associate each data point with each cluster using membership function*

- Fuzzy k-means
- Fuzzy k-modes
- The c-means Method

Centre-based: *Cluster have convex shapes and each cluster is represented by a centre. Its tries to minimize its objective function.*

- k-means
- k-Medoids
- k-modes

Graph- based clustering

{Clustering a set of graphs & clustering nodes/edges of a single graph}

Used in image segmentation and complex network analysis

- Chameleon
- CACTUS
- ROCK

Grid- based Clustering

{Create a grid structure, and the comparison is performed on grids}

Used for a multi-dimensional data set.

- CLIQUE
- STING
- CLARANS
- BIRCH

Density- based Clustering

{These clusters are areas of higher density than the remainder of the data set}

Uses SCAN method and requires density parameter as termination condition. It is used to identify clusters of arbitrary size and manage noise in data clusters

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- BRIDGE (integrates k-means and DBSCAN)

Model- based Clustering

{Algorithms designed for modelling unknown distribution as a mixture of simpler distributions}

Assumes that the data were generated by a model and tries to recover the original model from the data

- Dirichlet
- COOLCAT
- STUCCO

Clustering Analysis (R code)

K means (X, C, N)

X: data matrix
C: distinct number of cluster centers
N: seed value