UNIT-III

Regression – Concepts, Blue property assumptions, Least Square Estimation, Variable Rationalization, and Model Building etc. Logistic Regression: Model Theory, Model fit Statistics, Model Construction, Analytics applications to various Business Domains etc.

Regression:

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor).

This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Regression analysis is an important tool for modelling and analysing data

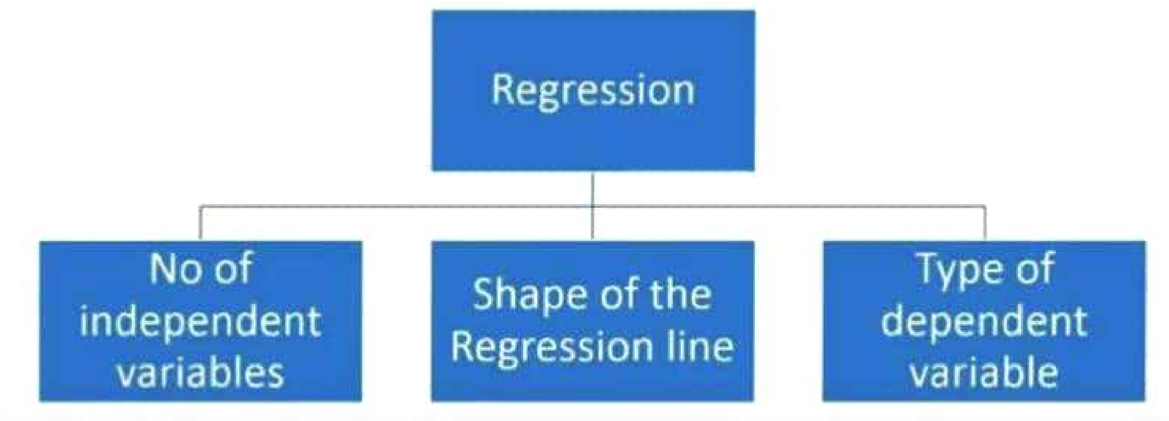
Why do we use Regression Analysis?

There are multiple benefits of using regression analysis. They are as follows:

- It indicates the significant relationships between dependent variable and independent variable.
- It indicates the strength of impact of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line).


Regression
No of independent variables
Shape of the Regression line
Type of dependent variable

1. Linear Regression

It is one of the most widely known modelling techniques. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).

It is represented by an equation Y=a+b*X + e, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable. Now, the question is "How do we obtain best fit line?".

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. 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 the line.

Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

Important Points:

- There must be linear relationship between independent and dependent variables
- Multiple regressions suffer from multicollinearity, autocorrelation, heteroskedasticity.
- Linear Regression is very sensitive to Outliers. It can terribly affect the regression line and eventually the forecasted values.
- Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
- In case of multiple independent variables, we can go with forward selection, backward elimination and step wise approach for selection of most significant independent variables.

2. Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is "why have we used log in the equation?"

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is logit function

Important Points:

- Logistic regression is widely used for classification problems
- Logistic regression doesn't require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
- To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression
- It requires large sample sizes because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
- The independent variables should not be correlated with each other i.e. no multi collinearity. However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
- If the values of dependent variable is ordinal, then it is called as Ordinal logistic regression
- If dependent variable is multi class then it is known as Multinomial Logistic regression.

3. Polynomial Regression

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

$$y=a+b*x^2$$

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

Important Points:

- While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem

4. Stepwise Regression

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves no human intervention.

This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping co-variates one at a time based on a specified criterion. Some of the most commonly used Stepwise regression methods are listed below:

- Standard stepwise regression does two things. It adds and removes predictors as needed for each step.
- Forward selection starts with most significant predictor in the model and adds variable for each step.
- Backward elimination starts with all predictors in the model and removes the least significant variable for each step.

The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle higher dimensionality of data set.

5. Ridge Regression

Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Above, we saw the equation for linear regression. Remember? It can be represented as:

y=a+ b*x

This equation also has an error term. The complete equation becomes:

y=a+b*x+e (error term), [error term is the value needed to correct for a prediction

error between the observed and predicted value]

=> y=a+y= a+ b1x1+ b2x2+....+e, for multiple independent variables.

Ridge regression solves the multicollinearity problem through shrinkage parameter λ (lambda).

Important Points:

- The assumptions of this regression is same as least squared regression except normality is not to be assumed
- Ridge regression shrinks the value of coefficients but doesn't reaches zero, which suggests no feature selection feature

6. Lasso Regression

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. Look at the equation below: Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables.

Important Points:

- The assumptions of lasso regression is same as least squared regression except normality is not to be assumed
- Lasso Regression shrinks coefficients to zero (exactly zero), which certainly helps in feature selection
- Lasso is a regularization method and uses l1 regularization
- If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero

7. ElasticNet Regression

ElasticNet is hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.

A practical advantage of trading-off between Lasso and Ridge is that, it allows Elastic-Net to inherit some of Ridge's stability under rotation.

Important Points:

- It encourages group effect in case of highly correlated variables
- There are no limitations on the number of selected variables
- It can suffer with double shrinkage

8. ROBUST REGRESSION

In robust statistics, robust regression is a form of regression analysis designed to overcome some limitations of traditional parametric and non-parametric methods. Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable.

Applications:

Heteroscedastic errors

One instance in which robust estimation should be considered is when there is a strong suspicion of heteroscedasticity. In the homoscedastic model, it is assumed that the variance of the error term is constant for all values of x. Heteroscedasticity allows the variance to be dependent on x, which is more accurate for many real scenarios.

Presence of outliers

Another common situation in which robust estimation is used occurs when the data contain outliers. In the presence of outliers that do not come from the same data-generating process as the rest of the data, least squares estimation is inefficient and can be biased. Because the least squares predictions are dragged towards the outliers, and because the variance of the estimates is artificially inflated, the result is that outliers can be masked. (In many situations, including some areas of geostatistics and medical statistics, it is precisely the outliers that are of interest.)

BLUE PROPERTY ASSUMPTIONS

The Least Squares modelling procedure is the Best Linear Unbiased Estimator (i.e BLUE). The Simple regression model needs five fundamental assumptions to be satisfied and the multiple regression model needs six assumptions to be satisfied. Among these four assumptions are related to models residuals. They are as follows.

1. The Residuals are distributed normally with zero mean.

2. The residuals maintain constant variance i.e there is no heteroscedasticity.

3. The successive residuals are not correlated and there is no chance for auto correlation.

VARIBALE RATIONALIZATION

Variable Rationalization is process of clustering data sets into more manageable parts for optimizing the query performance. It is used to divide the data in different ways. This technique is mostly applied on tables with partitioning or with out portioning. It allows performing sampling and map side joins. It places the data into a set number of files in every partition. This process can be assumed as grouping of objects by attributes.

Variable Rationalization is a method that increases the performance of big data operations. The data must be understood well before applying variable rationalization. The data might be complex and difficult to be understood. In such cases trail and error enable the users to better understand the data distribution or even guide them to follow right path. With Variable rationalization it is possible to create multiple and small parts based on the column values of the table.

Variable Rationalization is different from partitioning where every partition contains segments of files. More over partitioning can lead to deep and smaller partitions and directories which maximize the number of files. It even reduces the Name node performance and increases overhead. Variable rationalization generates the files in leaf-level directories that indicate the records with same column value.

Advantages:

- It generates faster responses for queries such partitioning
- Joins at Map side are quicker because of equal volumes of data in every partition.
- Improved Performance
- Provides tools to improve the performance of Bigdata operations.

Disadvantages:

- Programmers need to manually load equal amount of data.
- Programmers need to understand the data before applying the tools.
- Difficult to understand data in this process.

Model Building

In regression analysis, model building is the process of developing a probabilistic model that best describes the relationship between the **dependent and independent variables.** The major issues are finding the proper form (linear or curvilinear) of the relationship and selecting which independent variables to include. In building models it is often desirable to use qualitative as well as quantitative variables.

Quantitative variables measure how much or how many where as qualitative variables represent types or categories.

A regression analysis is typically conducted to obtain a model that may needed for one of the following reasons:

- To explore whether a hypothesis regarding the relationship between the response and predictors is true.

- To estimate a known theoretical relationship between the response and predictors. The model will then be used for:

- Prediction: The model will be used to predict the response variable from a chosen set of predictors, and

- Inference: The model will be used to explore the strength of the relationships between the response and the predictors

Therefore, steps in model building may be summarized as follows:

1. Choosing the predictor variables and response variable on which to collect the data.

2. Collecting data.
   You may be using data that already exists (retrospective), or you may be conducting an experiment during which you will collect data (prospective).

3. Exploring the data.
   - Check for data errors and missing values.
   - Study the bivariate relationships to reveal other outliers and influential Observations, relationships, and identify possible multicollinearity to suggest possible transformations.

4. Dividing the data into a model-building set and a model-validation set:
   - The training set is used to estimate the model.
   - The validation set is later used for cross-validation of the selected model.

5. Identify several candidate models:
   - Use best subsets regression.
   - Use stepwise regression.

6. Evaluate the selected models for violation of the model conditions. Below checks may be performed visually via residual plots as well as formal statistical tests.

- Check the linearity condition.
- Check for normality of the residuals.
- Check for constant variance of the residuals.
- After time-ordering your data (if appropriate), assess the independence of the Observations.
- Overall goodness-of-fit of the model.

7. Select the final model:

- Compare the competing models by cross-validating them against the validation data.

For Example suppose if we predict sales of an iced tea that is available in either bottles or cans. Clearly, the independent variable **"container type"** could influence the dependent variable **"sales."** Container type is a qualitative variable, however, and must be assigned numerical values if it is to be used in a regression study. So-called dummy variables are used to represent qualitative variables in regression analysis.

The dummy variable x could be used to represent container type by setting **$x = 0$ if the iced tea is packaged in a bottle and $x = 1$ if the iced tea is in a can**. If the beverage could be placed in glass bottles, plastic bottles, or cans, it would require two dummy variables to properly represent the qualitative variable container type. **In general, k - 1 dummy variables are needed to model the effect of a qualitative variable that may assume k values.**

The general linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + . . . + \beta_p x_p + \varepsilon$ can be used to model a wide variety of curvilinear relationships between dependent and independent variables. For instance, each of the independent variables could be a nonlinear function of other variables. Sometimes it is necessary to transform the dependent variable in order to build a satisfactory model. A logarithmic transformation is one of the more common types.

Least Square Estimation:

Ordinary Least Squares (OLS) is the most common estimation method for linear models.

As long as the model satisfies the OLS assumptions for linear regression, you can rest easy knowing that you're getting the best possible estimates.

Regression is a powerful analysis that can analyze multiple variables simultaneously to answer complex research questions. However, if you don't satisfy the OLS assumptions, you might not be able to trust the results.

What Does OLS Estimate and what are Good Estimates?

Regression analysis is like other inferential methodologies. Our goal is to draw a random sample from a population and use it to estimate the properties of that population.

In regression analysis, the coefficients in the regression equation are estimates of the actual population parameters. We want these coefficient estimates to be the best possible estimates!

Suppose you request an estimate—say for the cost of a service that you are considering. How would you define a reasonable estimate?

- The estimates should tend to be right on target. They should not be systematically too high or too low. In other words, they should be unbiased or correct on average.
- Recognizing that estimates are almost never exactly correct, you want to minimize the discrepancy between the estimated value and actual value. Large differences are bad!

The above two properties are exactly what we need for our coefficient estimates!

When your linear regression model satisfies the OLS assumptions, the procedure generates unbiased coefficient estimates that tend to be relatively close to the true population values (minimum variance). In fact, the Gauss-Markov theorem states that OLS produces estimates that are better than estimates from all other linear model estimation methods when the assumptions hold true.

The Seven Classical OLS Assumptions

Ordinary least squares (OLS) regression has underlying assumptions. When these classical assumptions for linear regression are true, ordinary least squares produces the best estimates. However, if some of these assumptions are not true, you might need to employ remedial measures or use other estimation methods to improve the results.

Many of these assumptions describe properties of the **error term**. The error term is a population value that we will never know. We will use the next best thing that is available—the residuals.

Residuals are the sample estimate of the error for each observation.

Residuals = Observed value – the fitted value

When it comes to checking OLS assumptions, assessing the residuals is crucial.

There are seven classical OLS assumptions for linear regression.

The first six are mandatory to produce the best estimates. While the quality of the estimates does not depend on the seventh assumption, analysts often evaluate it for other important reasons.

Assumption 1: The regression model is linear in the coefficients and the error term

This assumption addresses the functional form of the model. In statistics, a regression model is linear when all terms in the model are either the constant or a parameter multiplied by an independent variable. You build the model equation only by adding the terms together. These rules constrain the model to one type:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

In the equation, the betas (βs) are the parameters that OLS estimates. Epsilon (ε) is the random error.

Assumption 2: The error term has a population mean of zero

The error term accounts for the variation in the dependent variable that the independent variables do not explain. Random chance should determine the values of the error term. For your model to be unbiased, the average value of the error term must equal zero.

Assumption 3: All independent variables are uncorrelated with the error term

If an independent variable is correlated with the error term, we can use the independent variable to predict the error term, which violates the notion that the error term represents unpredictable random error. We need to find a way to incorporate that information into the regression model itself.

This assumption is also referred to as exogeneity. When this type of correlation exists, there is endogeneity. Violations of this assumption can occur because there is simultaneity between the independent and dependent variables, omitted variable bias, or measurement error in the independent variables.

Assumption 4: Observations of the error term are uncorrelated with each other

One observation of the error term should not predict the next observation. For instance, if the error for one observation is positive and that systematically increases the probability that the following error is positive, that is a positive correlation. If the subsequent error is more likely to have the opposite sign, that is a negative correlation. This problem is known both as serial correlation and autocorrelation.

Assess this assumption by graphing the residuals in the order that the data were collected. You want to see randomness in the plot. In the graph for a sales model, there appears to be a cyclical pattern with a positive correlation.

Assumption 5: The error term has a constant variance (no heteroscedasticity)

The variance of the errors should be consistent for all observations. In other words, the variance does not change for each observation or for a range of observations. This preferred condition is known as homoscedasticity (same scatter). If the variance changes, we refer to that as heteroscedasticity (different scatter).

The easiest way to check this assumption is to create residuals versus fitted value plot. On this type of graph, heteroscedasticity appears as a cone shape where the spread of the residuals increases in one direction. In the graph below, the spread of the residuals increases as the fitted value increases.

Assumption 6: No independent variable is a perfect linear function of other explanatory variables

Perfect correlation occurs when two variables have a Pearson's correlation coefficient of +1 or -1. When one of the variables changes, the other variable also changes by a completely fixed proportion. The two variables move in unison.

Perfect correlation suggests that two variables are different forms of the same variable. For example, games won and games lost have a perfect negative correlation (-1). The temperatures in Fahrenheit and Celsius have a perfect positive correlation (+1).

OLS Assumption 7: The error term is normally distributed

OLS does not require that the error term follows a normal distribution to produce unbiased estimates with the minimum variance. However, satisfying this assumption allows you to perform statistical hypothesis testing and generate reliable confidence intervals and prediction intervals.

The easiest way to determine whether the residuals follow a normal distribution is to assess a normal probability plot. If the residuals follow the straight line on this type of graph, they are normally distributed.

LOGISTIC REGRESSION:

Logistic regression is a transformation of the linear regression model that allows us to **probabilistically model binary variables.** It is also known as a generalized linear model that uses a logit-link.

We have to **use** logistic regression:

When you want to model binary data: Logistic regression is a go-to model for this use case. It models the probability that an observation takes on one of those two values. (Note: The model only predicts a probability, not class. Your choice of decision threshold depends on use case.)

When you want class probability predictions: Rather than class predictions alone, as in SVMs. This lets you gauge the confidence of your model in its predictions, and to play around with class decision thresholds.

When you want an interpretable model: Logistic regression is an ideal candidate model when you need to be able to explain the impact of each predictor. A predictor's coefficients quantify the impact of each feature on your model's predictions via the odds ratio.

When the decision boundary is smooth and linear: Logistic regression draws a smooth, linear decision boundary between two classes. Thus if your classes are linearly separable (you can separate points in n-dimensional space using *n - 1* dimensions), logistic regression will perform very well. You can also test for linear separability using a linear support vector machine.

We should **not use** the Logistic Regression:

When your data is not linearly separable: If you believe this is the case, consider using support vector machines with complex kernels or tree-based methods for classification.

When your goal is primarily performance: Logistic regression is a relatively simple model and will generally underperform against more complex models. If your goal is primarily prediction accuracy rather than model interpretation, boosted trees or neural nets may be a better choice.

Why do we use logistic regression?

We use logistic regression because linear regression is not appropriate for modeling binary outcomes. The below are the two reasons

A linear model makes continuous predictions that are unbounded. In binary classification, **we are interested in the probability of an outcome occurring, so we want predictions that are bounded between 0 and 1.**

Predicting binary outcomes with a linear model violates the assumption of normal residuals, distorting inferences made on regression coefficients.

MODEL CONSTRUCTION

Data Modelling is the process of creating a data model for storing the data in the Data base. A model is nothing but representation of data objects that is associated between different data objects and rules. Data modelling helps in visually representing the data and the enforce business rules, regulatory compliances and the government policies on data. It assures consistency in naming conventions, semantics, default values and security while assuming quality. The Data model focuses on the data required. It even focuses on the operations that are to be performed on the data. The data model enables to build the conceptual model and to set the relationship between the data items. The Mostly used data modelling techniques are Entity Relation Ship Model (E-R) and Unified Modeling Language (UML).

Data Modeling is a complex science needs the logical relationships to be designed to interrelate the data with each other. It even supports the Business.

The logical designs are translated into physical models which contain storage devices, databases and the files that build the data. The businesses earlier used the relation data base technology such as SQL to build the data models.

Data Analytics consists of large percentage of data under management and it does not run on relational databases. It runs on non-relational data bases such as NoSQL.

An efficient model can be constructed by following the steps

1. Do not Impose Traditional Modelling Techniques on Data

Fixed record data is stable and even predictable in its growth. With this data modelling become easy and effort of modelling must center on building open and elastic data interfaces.

2. Design a System rather Schema

In Relational Database schema the relationships and links between the data needed by business for its information support. In case of Bigdata it doesn't have a database or it uses database NoSQL. The Bigdata models must be created on systems rather than on databases.

3. Use Data Modelling Techniques

The IT Decision makers must include the ability to create the data models for big data as the requirements while considering the Bigdata tools and methodologies.

4. Focus on the Core Data of Business

Enterprises get the data at large volumes. Most of the data is irrelevant. The Best method would be to identify the Bigdata suitable for enterprise and to model it.

5. Deliver the Quality of Data

Earlier data models and relationships are affected for big data when organizations focus on development of sound definitions for data. The knowledge about the data helps in placing it properly in data models to support business.

6. Search for Key in Roads into the Data.

A Commonly used vector into big data today is geographical location. The data models can be created by supporting information access paths for the company identifying the common entry points into the data.

ANALYTICS APPLICATIONS TO VARIOUS BUSINESS DOMAINS

Some of the different data analytics applications that are currently being used in several organizations across the globe are:

1. Security

Data analytics applications or, more specifically, predictive analysis has also helped in dropping crime rates in certain areas. In a few major cities like Los Angeles and Chicago, historical and geographical data has been used to isolate specific areas where crime rates could surge. On that basis, while arrests could not be made on a whim, police patrols could be increased. Thus, using applications of data analytics, crime rates dropped in these areas.

2. Transportation

Data analytics can be used to revolutionize transportation. It can be used especially in areas where you need to transport a large number of people to a specific area and require seamless transportation. This data analytical technique was applied in the London Olympics a few years ago.

For this event, around 18 million journeys had to be made. So, the train operators and TFL were able to use data from similar events, predict the number of people who would travel, and then ensure that the transportation was kept smooth.

3. Risk detection

One of the first data analytics applications may have been in the discovery of fraud. Many organizations were struggling under debt, and they wanted a solution to this problem. They already had enough customer data in their hands, and so, they applied data analytics. They used 'divide and conquer' policy with the data, analyzing recent expenditure, profiles, and any other important information to understand any probability of a customer defaulting. Eventually, it led to lower risks and fraud.

4. Risk Management

Risk management is an essential aspect in the world of insurance. While a person is being insured, there is a lot of data analytics that goes on during the process. The risk involved while insuring the person is based on several data like actuarial data and claims data, and the analysis of them helps insurance companies to realize the risk.

Underwriters generally do this evaluation, but with the advent of data analysis, analytical software can be used to detect risky claims and push such claims before the authorities for further analysis.

5. Delivery

Several top logistic companies like DHL and FedEx are using data analysis to examine collected data and improve their overall efficiency. Using data analytics applications, the companies were able to find the best shipping routes, delivery time, as well as the most cost-efficient transport means. Using GPS and accumulating data from the GPS gives them a huge advantage in data analytics.

6. Fast internet allocation

While it might seem that allocating fast internet in every area makes a city 'Smart', in reality, it is more important to engage in smart allocation. This smart allocation would mean understanding how bandwidth is being used in specific areas and for the right cause.

It is also important to shift the data allocation based on timing and priority. It is assumed that financial and commercial areas require the most bandwidth during weekdays, while residential areas require it during the weekends. But the situation is much more complex. Data analytics can solve it.

For example, using applications of data analysis, a community can draw the attention of high-tech industries and in such cases, higher bandwidth will be required in such areas.

7. Reasonable Expenditure

When one is building Smart cities, it becomes difficult to plan it out in the right way. Remodelling of the landmark or making any change would incur large amounts of expenditure, which might eventually turn out to be a waste. Data analytics can be used in such cases. With data analytics, it will become easier to direct the tax money in a cost-efficient way to build the right infrastructure and reduce expenditure.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

8. Interaction with customers

In insurance, there should be a healthy relationship between the claims handlers and customers. Hence, to improve their services, many insurance companies often use customer surveys to collect data. Since insurance companies target a diverse group of people, each demographic has their own preference when it comes to communication.

Data analysis can help in zeroing in on specific preferences. For example, a study showed that modern customers prefer communication through social media or online channels, while the older demographic prefers telephonic communication.

9. Planning of cities

One of the untapped disciplines where data analysis can really grow is city planning. While many city planners might be hesitant towards using data analysis in their favour, it only results in faulty cities riddled congestion. Using data analysis would help in bettering accessibility and minimizing overloading in the city.

Overall, it will generate more efficiency in the planning process. Just erecting a building in a suitable spot will not create an overall benefit for a city since it can harm the neighbours or the traffic in the area. Using data analytics and modelling, it will be easy to predict the outcome of placing a building in a specific situation and therefore, plan accordingly.

10. Healthcare

While medicine has come a long way since ancient times and is ever-improving, it remains a costly affair. Many hospitals are struggling with the cost pressures that modern healthcare has come with, which includes the use of sophisticated machinery, medicines, etc.

But now, with the help of data analytics applications, healthcare facilities can track the treatment of patients and patient flow as well as how equipment are being used in hospitals. It has been estimated that there can be a 1% efficiency gain achieved if data analytics became an integral part of healthcare, which will translate to more than $63 billion in healthcare services.

11. Travelling

If you ever thought travelling is a hassle, then data analytics is here to save you. Data analysis can use data that shows the desires and preferences of different customers from social media and helps in optimizing the buying experience of travellers. It will also help companies customize their own packages and offer and hence boost more personalized travel recommendations with the help data collected from social media.

12. Managing Energy

Many firms engaging with energy management are making use of applications of data analytics to help them in areas like smart-grid management, optimization of energy, energy distribution, and automation building for other utility-based companies. How does data analytics help here?

Well, it helps by focusing on controlling and monitoring of a dispatch crew, network devices, and management of service outages. Since utilities integrate about millions of data points within the network performance, engineers can use data analytics to help them monitor the entire network.

13. Internet searching

When you use Google, you are using one of their many data analytics applications employed by the company. Most search engines like Google, Bing, Yahoo, AOL, Duckduckgo, etc. use data analytics. These search engines use different algorithms to deliver the best result for a search query, and they do so within a few milliseconds. Google is said to process about 20 petabytes of data every day.

14. Digital advertisement

Data analytics has revolutionized digital advertising, as well. Digital billboards in cities as well as banners on websites, that is, most of the advertisement sources nowadays use data analytics using data algorithms. It is one of the reasons why digital advertisements are getting more CTRs than traditional advertising techniques. The target of digital advertising nowadays is focused on the analysis of the past behaviour of the user.

It is clear that data analytics applications are taking great strides in almost all avenues across the globe. If we are able to understand data and analyze it, it can help in increasing our overall job efficiency a lot. However, misuse or inefficient use of data can cause several problems and lead to the lowering of overall productivity.

Model Theory

Model theory is one of the branches of mathematical logic that deals with abstract data structures.

Model theory is divided into two parts

1. Pure

2. Applied

Pure model theory will learn the abstract properties of first order theories and there on derives structure theorem and models.

Applied model theory will study the concrete algebraic structures from model theoretic point of view and then uses the results from pure model theory functionalities and uniformities of definition. The applied model theory is connected strongly with other branch of mathematics. The results if it will have non-model theoretic implications.

The other areas of model theory are

1. Pure Model Theory

The further development of model theory are expected in usage of stability theory technique in unstable contexts.

2. Model Theory of Fields with Operators and Connections with Arthimetic Geometry

3. Henselian Fields

These are concerned with model theory connections with algebraic and analytic geometry.

4. O-Minimality and Related Topics.

Property of Ordered Structures which generates results related to traditional real analytic results.

Differences between Correlation and Regression

Correlation	Regression
Correlation describes the degree or strength of relationship between two or more variables.	Regression gives the average relationship if it exists between two or more variables
It determines whether a relationship exists between two variables or not	If there exists a relationship between two variables then it predicts or estimates the value of the dependent variable for any given value of the independent variable.
Correlation coefficient between two variables X and Y is symmetric i.e r(X,Y)=r(Y,X)	Regression coefficient of X on Y b_{xy} is not symmetrical to regression coefficient of Y on X i.e $b_{xy} \neq b_{yx}$

Model Fit Statistics

Model fit Statistics describe and test the overall fit of the model.

The model fit measures the similarity between fitted model and actual outcome values that are generated. Model misspecification will indicate the selection of variables and recording or transformation of them and how they fit model to best approximate. The model misspecification tests are developed for examining several aspects of model misspecification in contrast to the typically used summary level goodness of fit tests. The model fit as well as model misspecification capture various aspects of model validity. With respect to data interpretation model fit is a measure of discrepancy in between the observed empirical distribution of observation in data set and best fitting probability distribution that calculated from estimated probability model. The model parameters might be estimated to fit the model in the presence of parameterized model and data.

The problem of assessing model fit might be challenging if researchers tend to measure the fit that accounts for variability in model complexity, model misspecification and small sample size.

The examples of global or summary model fit measures which are mostly used to assess overall model fit are Sum of Squared Errors(SSE) and Log Likelihood(LL) and model selection criteria.

S.No.	Model Fit Measures	Description
1.	Sum of Squared Errors (SSE)	Sum of squared differences between predicted and observed values. Measures deviation from actual values.
2.	R^2, adjusted $R^{2'}$, Pseudo $R^{2'}$ statistics	The coefficients of determination (R^2) compares the predictive performance of model with constrained version of model.
3.	Log-likelihood (LL)	The Kullback-Leibler based measure of model fit to observed data. It will choose the model which can most likely create the in sample data.