ANALYTIXLABS

Introduction to Predictive Modelling and Linear & Logistic Regression

What are the Typical Business problems which we encounter?

What are the Typical Business problems?

- How to attract new customers?
- How to make those new customer to be profitable?
- How to avoid high risk/bad customers?
- How to understand the characteristics of current customers?
- How to make your unprofitable customers more profitable?
- How to retain your profitable customers?
- How to win back your existing customers?
- How to improve customer satisfaction?
- How to increase sales/profit and reduce expenses?
- How to recommend products to customers?
- How to optimize marketing expenses?
- How to take decision for offering credit card?
- How to increase credit line for given customer?
- How to optimize collection process?
- How to detect fraud transaction/customer?
- How to price the product?
- How to identify visitor will click or not?
- How to identify to employees who attrite?
- How to identify when customer stops buying/using product?
- How to predict how much customer make purchase?
- How to predict how much loss given the customer stop using product?
- how to calculate the impact of sales/volume given the price change?
- How to forecast the sales for next two quarters?
- How to optimize cash flows and funds utilization?
- How to optimize cash in ATMS?
- Does income of individual depend on demographics (Age and Years of education) and others?
- Which of the retail image levers drives footfalls or conversions?
- What drives satisfaction among branch users?
- What causes high performance of bank branch on the basis of financial parameters?

ANALYTIXLABS

Lets deep dive some of the problems!

ANALYTIXLABS

Example

In a credit card business. Applications have come for new card, bank has to take decision on whether to approve the credit or not and decide how much credit line need to be granted for given application?

Question

- Should we grant him/her the card?
- how much credit line need to be offered?

Non-deterministic information (Y)

- Chances that the customer will default on his/her payments
- The maximum amount ($) that we may approve

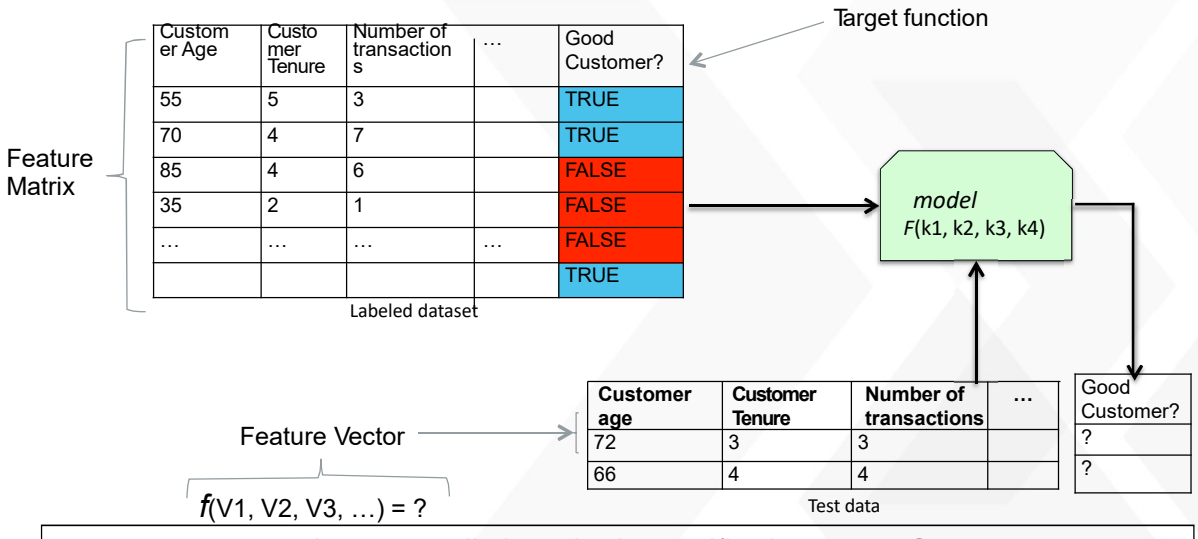
Known information (X)

- Information on credit history, past transactions, financial status of the customer.

A Functional relationship between X and Y helps deciding whether to approve the credit request

ANALYTIXLABS

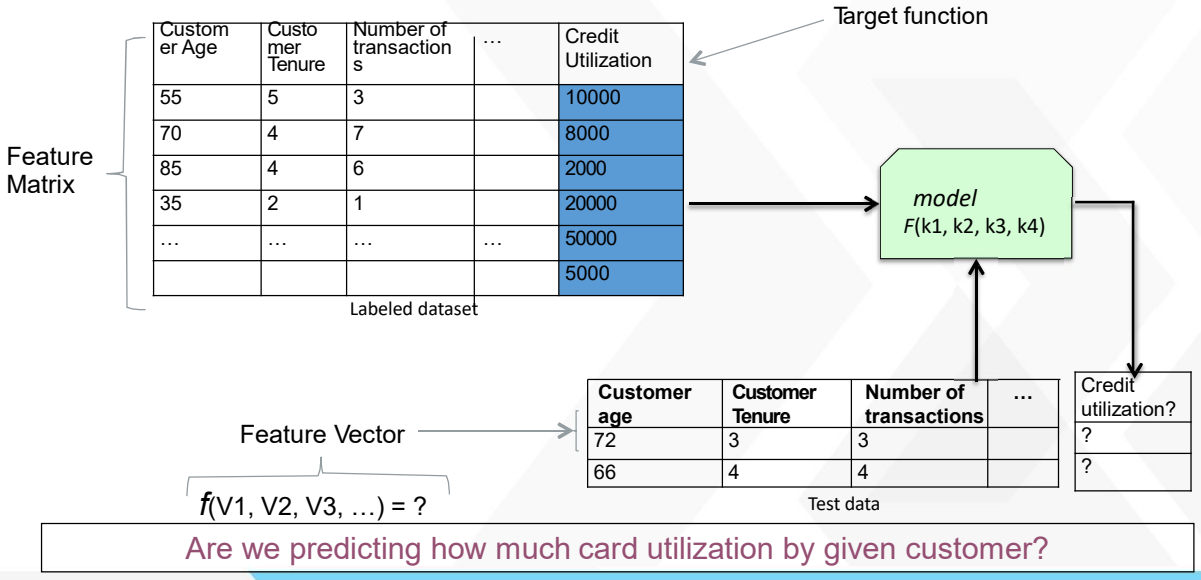
Example- should we grant him credit card?



Are we predicting who is good/bad customer?

ANALYTIXLABS

Example – How much credit line need to be offered?



ANALYTIXLABS

How to classify these problems?

Business problems – Types:

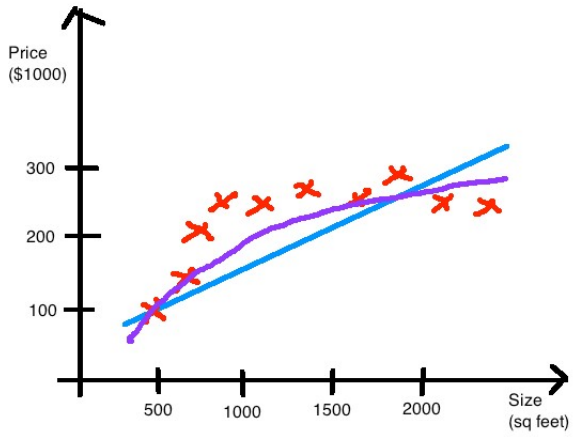
- Regression Problems – predicting a value
- Classification problems – predicting an event or predicting a category
- Segmentation problems – classifying the data when we don't have target variable(Un-supervised classification)
- Forecasting problems – Predicting future value(It is similar to regression however one of the independent variable is time)
- Others – rest of all like optimization problems, survival problems etc…

ANALYTIXLABS

Regression Problems

ANALYTIXLABS

Regression: predict a continuous value

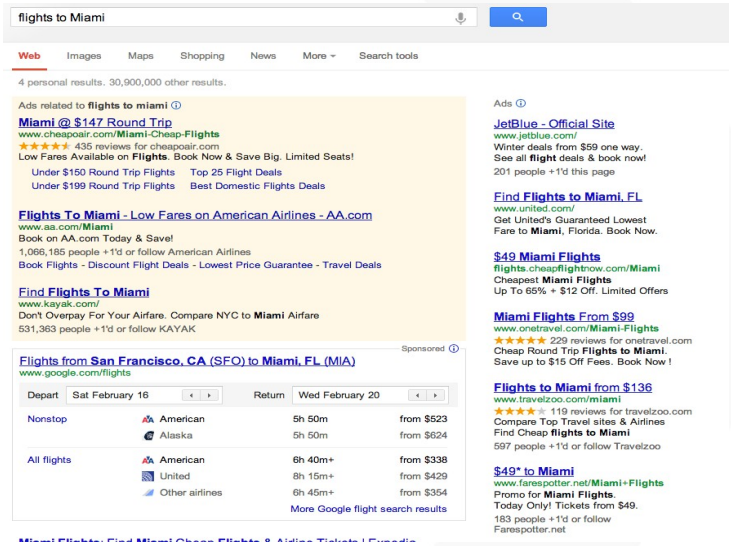


Some techniques:

- Linear Regression / GLM
- Decision Trees
- Support vector regression
- SGD
- Ensembles

ANALYTIXLABS

Regression Example: Ad Click-Through Rates in Ad Search



Rank = bid * CTR

Predict CTR for each ad to determine placement, based on:

- Historical CTR
- Keyword match
- Etc…

ANALYTIXLABS

Regression – Typical Applications

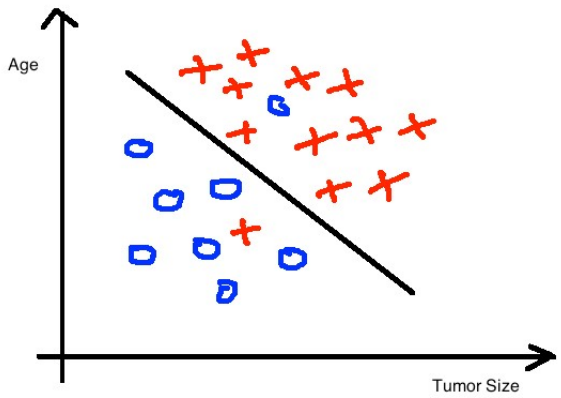
- Typical Applications:
  - Stock market: predict the share price for the future
  - Does income of individual depend on demographics (Age and Years of education) and others?
  - Which of the retail image levers drives footfalls or conversions?
  - What drives satisfaction among branch users?
  - What causes high performance of bank branch on the basis of financial parameters?
  - Energy demanding in a dam
  - Wind speed: eolic energy
  - Travel time prediction: for the planning of transport companies
  - Level of water in a river: for safety & prevention
  - Tax income: public budget
  - …

ANALYTIXLABS

Classification Problems

ANALYTIXLABS

Classification: predicting a category



Some techniques:

- Naïve Bayes
- Decision Tree
- Logistic Regression
- SGD
- Support Vector Machines
- Neural Network
- Ensembles

ANALYTIXLABS

Detailed list of classification Techniques

Classical Techniques

- Logistic Regression
- Decision Trees (CHAID/CART)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis(QDA)

Ensemble Learning

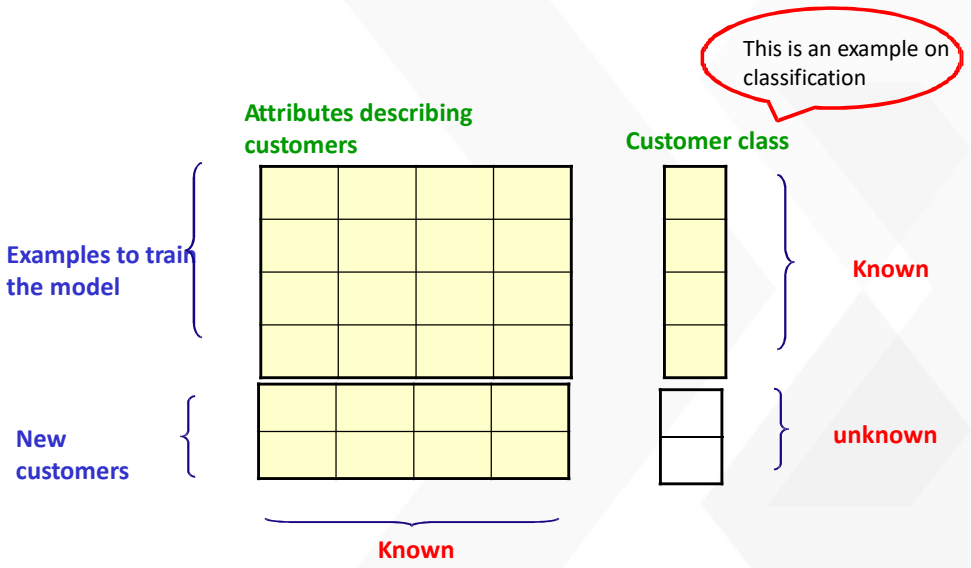
- Bootstrapped Aggregation(Bagging)
- Boosting (AdaBoost/Gradient Boosting Machines)
- Random Forecast

Machine Learning Techniques

- K-Nearest Neighbours (KNN)
- Naïve Bayes
- Artificial Neural Networks (ANN)
- Support Vector Machines (SVM)

ANALYTIXLABS

Classification Example



ANALYTIXLABS

Classification – Typical applications

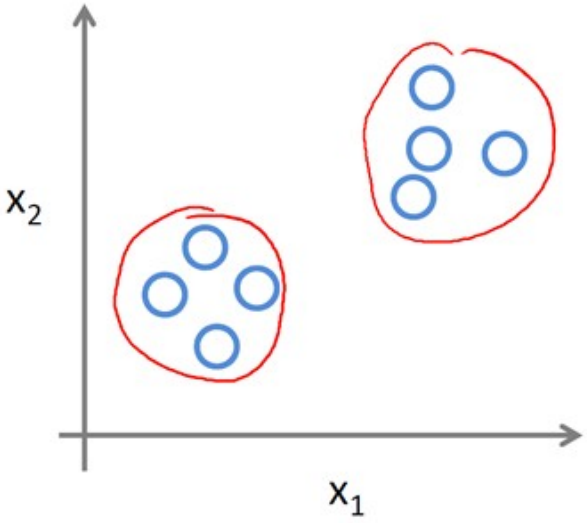
- Typical Applications:
  - Credit approval: classifies credit application as low risk, high risk, or average risk
  - Determine if a local access on a computer is legal or illegal
  - Target marketing (send or not a catalogue?)
  - Fraud Detection: Fraud Vs. Not Fraud
  - Collections: Identify cardholders that are likely to default and thus need collection effort (Payment Projection Models)
  - Insurance: Identify claims that are Fraud or Not Fraud
  - Marketing & Sales: Identify to responders to promotional campaigns(Response/Non Response, Buying/Not Buying
  - Operations: Models to identify to employees who attrite(Attrition/ Retention)
  - Website: Models to identify to weather visitor will click or not(Click/Not Click)
  - Gaming: Models to identify to who will win(Win/Loss)
  - Health Care: Models to identify to cure or not cure(Cure/ Not Cure)
  - Text classification (spam, not spam)
  - Text recognition (Optical character recognition)
  - …

ANALYTIXLABS

Segmentation Problems

ANALYTIXLABS

Segmentation: detect similar instance groupings / detect natural patterns



Some techniques:
- k-means
- Hierarchical clustering
- Spectral clustering
- DB-scan

ANALYTIXLABS

Segmentation Example: Market Segmentation

Age	State	Annual Income	Marital status
25	CA	$80,000	M
45	NY	$150,000	D
55	WA	$100,500	M
18	TX	$85,000	S
…	…	…	…

No labels

Model

Naturally occurring (hidden) structure

ANALYTIXLABS

Example: market segmentation

ANALYTIXLABS

Segmentation – Typical applications

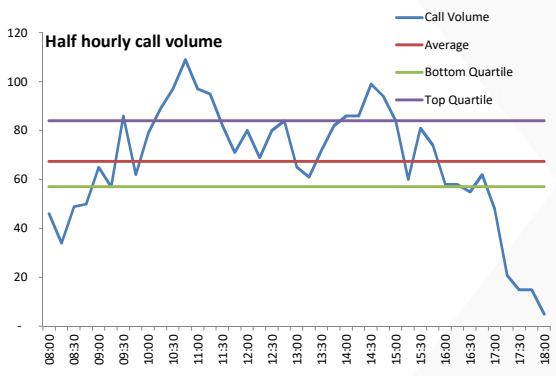
- Typical Applications:
  - Improve customer retention by providing products tailored for specific segments
  - Increase profits by leveraging disposable incomes and willingness to spend
  - Grow you business quicker by focusing marketing campaigns on segments with higher propensity to buy
  - Improve customer lifetime value by identifying purchasing patterns and targeting customers when they are in the market
  - Retain customers by appearing as relevant and responsive
  - Identify new product opportunities and improve the products you already have
  - Optimize operations by focusing on geographies, age groups etc. with the most value
  - Increase sales by offering free shipping to high frequency buyers
  - Offer improved customer support to VIP customers
  - Gain brand evangelists by incentivising them to comment, review or talk about your product with free gifts or discounts
  - Reactivate customers who have churned and no longer interact with you
  - …

ANALYTIXLABS

Forecasting Problems

ANALYTIXLABS

Forecasting: predict a continuous value for future(eg: next two quarters)



Some techniques:

- Averages
- Smoothening
- Decomposition
- ARIMA/SARIMA
- ARIMAX
- ARCH/GARCH
- VAR

etc…

ANALYTIXLABS

Forecasting – Typical applications

- Typical Applications:
  - Call volume demand in call centers
  - Average handle time trends
  - Demand for seasonal maintenance
  - Event based demand for field services
  - Estimation of cash requirement in ATMs and Branches
  - Number of transactions for tellers
  - Footfall estimation in consumer retail
  - IT manpower requirement over months
  - …

ANALYTIXLABS

Other Problems

ANALYTIXLABS

Example: Affinity Analysis- identifying frequent item sets

	Item 1	Item 2	Item 3	Item 4	Item 5	…
Tx 1	Y	N	N	Y	N	
Tx 2	Y	N	N	Y	N	
Tx 3	Y	Y	N	Y	N	
Tx 4	N	N	Y	Y	Y	
Tx 5						
…						

	Item 1	Item 2	Item 3	Item 4	Item 5	…
Tx 1	Y	N	N	Y	N	
Tx 2	Y	N	N	Y	N	
Tx 3	Y	Y	N	Y	N	
Tx 4	N	N	Y	Y	Y	
Tx 5						
…						

Goal: identify frequent item set
Techniques: FP Growth, a priori

ANALYTIXLABS

Example: Affinity Analysis

Use affinity analysis for
- store layout design
- Coupons

ANALYTIXLABS

Predictive Modeling

ANALYTIXLABS

What is Modeling?

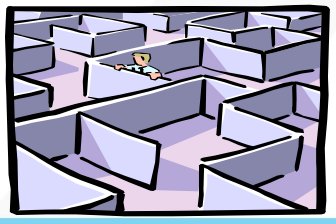
By "Modeling" we mean developing set of equations or mathematical formulation by which we can

- Predict certain events
- Identify the drivers of certain events based on some explanatory variables

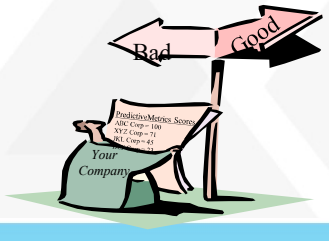
For example, we can build models to predict drivers of sales, risk of a borrower.

Historical Data

Statistical Analyses



Predict Future Events



ANALYTIXLABS

What is predictive Modelling?

Predictive Model:
Goal: To predict the value of a given variable (named **target** or **objective** variable)

$$y = f(\mathbf{x})$$

- Training: given a training set of labeled examples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set
- Testing: apply f to a never before seen test example x and output the predicted value **y = f(x)**

- For each record on the dataset determines the value of the class attribute
- Constructs a model based on the training set; then, uses the model in predicting new data

ANALYTIXLABS

Why Do We Estimate f?

Predictive Modeling is all about how to estimate f.

Why do we care about estimating f?

There are 2 reasons for estimating f,
- ✓Prediction
- ✓Inference

ANALYTIXLABS

Predictive Models: Examples

- Prospecting/Response model (stage 1): predict potential customers' likelihood of conversion
- Xsell model (stage 1): predict current customers' likelihood to purchase other products
- Balance model (stage 2): predict customers' opening balance if they open accounts
- Potential value model: measure customers' future (profit) potential
- Risk/Credit model: predict people's likelihood of default/charge off

Why do we need predictive Model?

- Distinguish/understand different types of customers in term of risk, potential value, likelihood of conversion/xsell/attrition, etc.
- Model helps us better targeting audience – higher conversion rate, lower marketing cost
- Compare with human judgment, model costs less and is more consistent, robust, efficient – easy to implement on large population
- With new information, model can be systematically evaluated and improved to enhance targeting

Overview of modeling key concept

Term	Examples	Description
Target / Y /Dependent Variable	Paid users among all the users	Represents the output or effect
X / Independent Variable	Age, gender, product usage	Represent the inputs or causes
Probability	26% to become paid user	Is a measure of the expectation that an event will occur or a statement is true
Score	0.26	Is the value of a variable below which a certain percent of observations fall
Percentile	75%	Is the value of a variable below which a certain percent of observations fall

ANALYTIXLABS

Define target variables

- **Look-a-like Model:** Use customer who are currently having the product as modeling target
- **Walk-in Model**: There is a modeling window. Customer who opened the product during this time period is defined as modeling target
- **Response Model**: Use customer who converted as a result of campaign as modeling target
- **Uplift Model**: Use the change in behavior as a result of a treatment as modeling target

	Pros	Cons
Look-a-like	When there are not enough modeling targets, look-a-like model is best way to remedy sample size issue.	Model works like a profile. It uses the differences between product holder and non-holder as main drivers. It could be misleading in cause-effect and event sequencing
Walk-in	There is a time window. Model captures the natural response. It's a good start when no campaign was ever launched.	It's still a retrospective model, not campaign driven. Does not capture marketing effect
Response	Uses the results of real campaigns. Natural response + marketing effect	Smaller sample size. Non-represetative sample of population - cut based on BAU and old models
Uplift	Identify the pursuable that will actually be influenced by your campaign, avoid targeting individuals that will buy anyway	No proven techniques yet to achieve reliable results as other modeling type

ANALYTIXLABS

Nature of Explanatory & Dependent variables

An Explanatory variable could be

✓ Numerical

Discrete : e.g. Number of satisfactory trades

Continuous : e.g. Highest Credit Line

✓ Categorical

Ordinal : e.g. Income Group (High/Medium/Low)

Nominal : e.g. Gender (Male/Female)

A Dependent variable could be

✓ Continuous: e.g. The total ($) that we may approve

✓ Discrete : e.g. Number of equipments that may be funded

✓ Binary : e.g. Whether the customer would default on payment or not (1/0)

ANALYTIXLABS

Analyze Data Major Steps

	Steps
0 Business Problem	▪ Convert business problem into statistical problem ▪ Identify type of problem - Technique ▪ Define Hypothetical relationship
1 Data Construction	▪ Create the model data by various sampling ▪ Aggregate the data at same level(eg: customer level) – Depends technique
2 Univariate/Bi-Variate Analysis	▪ Examine the data for errors, outliers and missing values. ▪ Assess/understand the relationship to target variable ▪ Understand the relationship between independent variables
3 Data Preparation	▪ Exclusions/Data type conversions/Outlier treatment/Missing value treatment ▪ Create new, hypothetically relevant variables, e.g. max, min, sum, change, ratio ▪ Binning variables – dummy variables creation ▪ Transform data to help ensure linearity
4 Variable Reduction	▪ Avoid collinearity and shorten computing time by reducing the number of independent variables – variable cluster, correlation, factor analysis etc.

ANALYTIXLABS

Variable reduction techniques

The following variable reduction techniques have been using as part of model development.

- ✓ Information Value or Weight-of-Evidence
- ✓ Principal Component Analysis /Factor Analysis
- ✓ Variable Classing (Variable clustering)
- ✓ Variance Inflation Factor(VIF) / Conditional Index(CI)
- ✓ Marginal Information Value
- ✓ Step-wise Variable Selection (Forward/Backward/stepwise)
- ✓ Univariate Analysis

ANALYTIXLABS

Model Development Major Steps

	Steps
1 Split the data for validation	▪ Split the file into the modeling(training) and validation(test) data sets
2 Build the Model on Training data	▪ Linear Regression/Logistic Regression/Decision Trees/Segmentation etc. ▪ Assumptions checks ▪ Modify the data as per the assumptions(if required)
3 Process the model	▪ Interpret the model ▪ Understand model diagnostics – accuracy(fit) of the model ▪ Iterate the models
4 Validate the model	▪ Re-run the model ▪ Using scoring ▪ Using cross validation(K-Fold) Validation
5 Implementation of the Model	▪ Prepare final model results– present the model ▪ Identify the limitations of the model ▪ Implement the model(converting stats solution into business solution)

ANALYTIXLABS

Regression Modeling

ANALYTIXLABS

Business Problem

I am the CEO of a hypermarket chain “Safegroceries” and I want to open new store which should give me the best sales . I am hiring “Alabs” to help me figure out a location where to open the new store

What should ALABS do ?

Additional Information about Safe groceries:

- Safegroceries has more than 5000 stores across the world
- It is upstream hypermarket store catering to high end products
- There are more than 100 locations he needs to choose from ?

ANALYTIXLABS

What could impact sales ?

- ✓ Population Density in the area
- ✓ Disposable Income
- ✓ Demographics of the region
- ✓ Parking size of the location
- ✓ No of other grocery stores in around (3km)
- ✓ Credit card usage
- ✓ Internet penetration/usage
- ✓ Average no of cars/household
- ✓ Avg family size/household
- ✓ No of working people/household
- ✓ ……..
- ✓ ……..

ANALYTIXLABS

Relationship between Sales and Variables

- ✓ Sales = function (X1, X2, X3, X4,X5,X6……)
- ✓ Sales = 10X1 + 20X2 +0.5X3 + 8X4 +……
- ✓ If the function is linear we call it linear regression

This was a case of prediction . How about doing root cause analysis ?

Now CEO wants to improve the performance of the existing stores and wants to increase sales ?

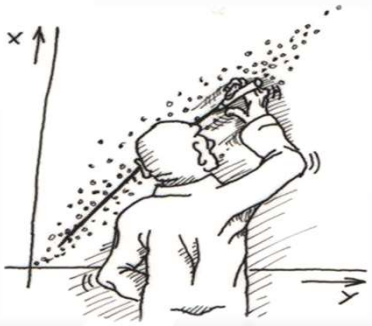
Decision – Prediction vs Inference(root causal)

ANALYTIXLABS

Regression

Regression Analysis

"Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another"



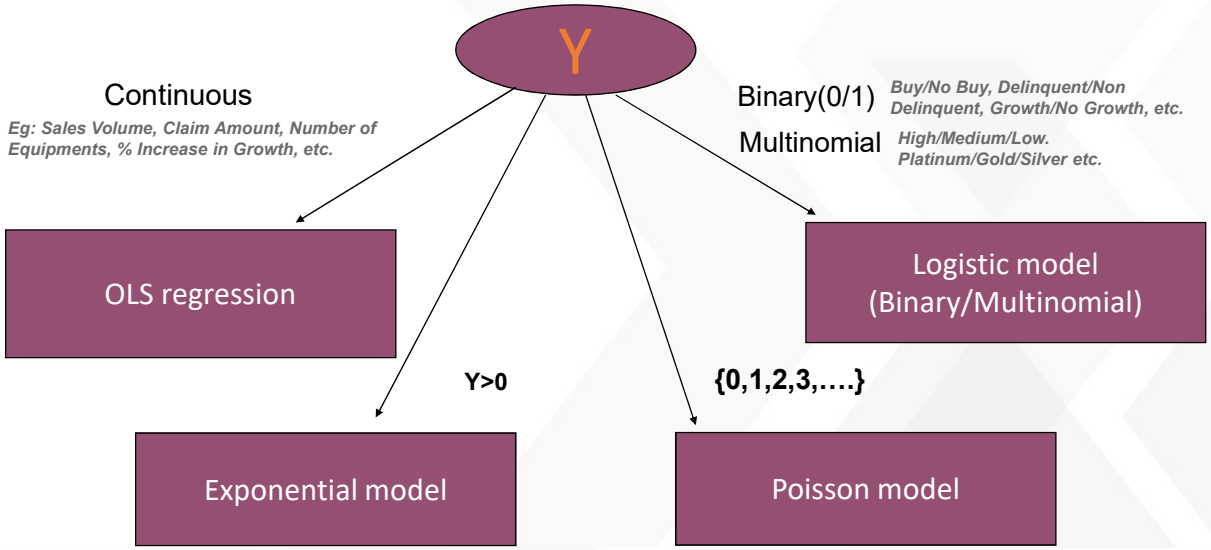
Regression modeling

Establishing a functional relationship between a set of Explanatory or Independent variables X_1, X_2, …, X_p with the Response or Dependent variable Y.

$$Y = f(X_1, X_2, .., X_p)$$

ANALYTIXLABS

Types of Regression Models



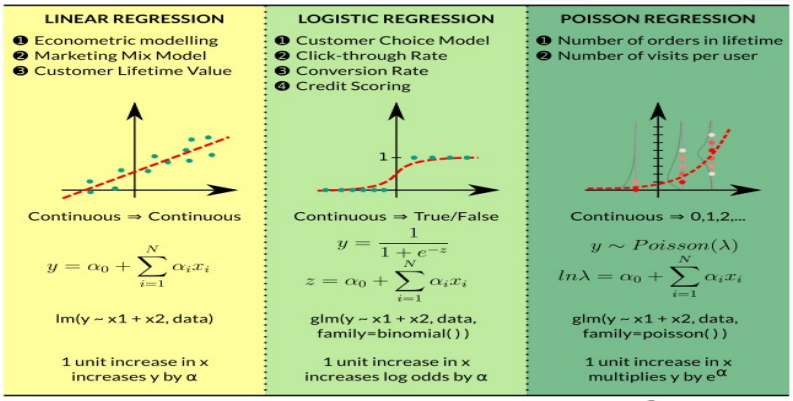
ANALYTIXLABS

Three Regression Types (GLM)

Generalized linear models extend the ordinary linear regression and allow the response variable y to have an error distribution other than the normal distribution.

GLMs are:

- A. Easy to understand
- B. Simple to fit and interpret in any statistical package
- C. Sufficient in a lot of practical applications



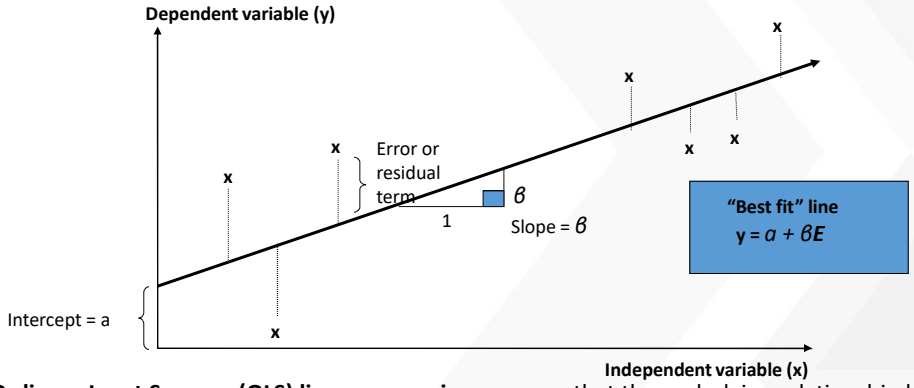
ANALYTIXLABS

Ordinary Least Square Regression(OLS)

ANALYTIXLABS

What is OLS REGRESSION ANALYSIS?

OLS Regression basically try to draw the best fit regression line - a line such that the sum of the squared deviations of the distances of all the points to the line is minimized.



Ordinary Least Squares (OLS) linear regression assumes that the underlying relationship between two variables can best be described by a line.

ANALYTIXLABS

Regression-Step-0

Step-0:

Identification of Dependent Variable
Example: Expected revenue from telecom license

Step-1:

Once we have selected the dependent variable we wish to predict, the first step before running a regression is to identify what independent variables might influence the magnitude of the dependent variable and why.

ANALYTIXLABS

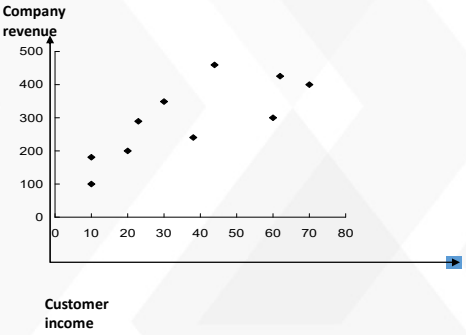
Regression-Step-1

COLLECTING AND GRAPHING THE DATA

The first step is to collect the necessary information and to enter it in a format that allows the user to graph and later "regress" the data.

(Y) Company revenue	(X) Customer income
180	10
100	10
200	20
290	23
350	30
240	38
460	44
300	60
425	62
400	70

Plotting the data allows us to get a "first look" at the strength of our relationship

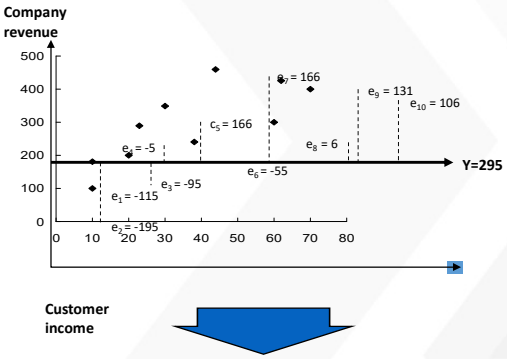


ANALYTIXLABS

Regression-Step-2

The way linear regression "works" is to start by naively fitting a horizontal no-slope (slope = A=0) line to the data. The y-intercept B of this line is simply the arithmetic average of the collected values of the dependent variable.

(Y) Company revenue	(X) Customer income
180	10
100	10
200	20
290	23
350	30
240	38
460	44
300	60
425	62
400	70
Average Y value = 295	



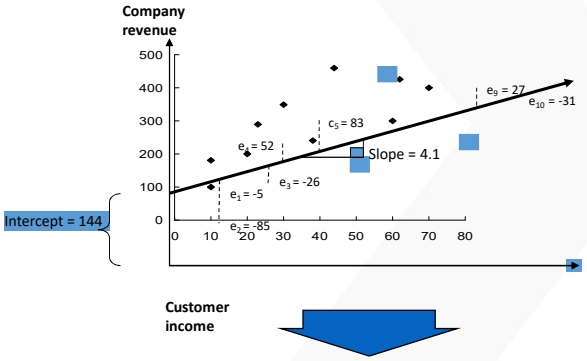
The sum of the squared residuals, $S_{no\text{-}slope}$ gives us a measure of how well the horizontal line fits the data

$$S_{no\text{-}slope} = (-115)^2 + (-195)^2 + (-95)^2 + (-5)^2 + \ldots (106)^2 = 121{,}523$$

ANALYTIXLABS

Regression-Step-3

If we allow the line to vary in slope and intercept, we should be able to find that line which minimizes the sum of squared residuals.



"Best fit" sloped line =

Revenue = 144 + 4.1 x (income)

The new sum of squared residuals, S_{slope}, should be lower than $S_{no\text{-}slope}$, if the new line provides a better fit to the data

$$S_{slope} = (-5)^2 + (-85)^2 + (-26)^2 + \ldots + (-31)^2 = 49{,}230$$

ANALYTIXLABS

Critical Elements of linear Regression

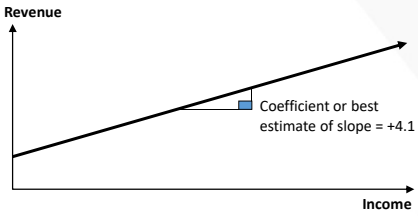
Since software packages like SAS/R will regress any stream of data regardless of its integrity, it is critical that we review the regression results first to determine if a meaningful relationship exists between the two variables before drawing any conclusions.

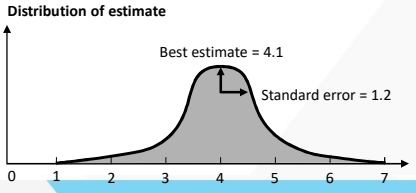
- **Sign and magnitude of coefficients**
- **T-statistics**
- **R2-statistics**

ANALYTIXLABS

Interpreting the coefficient – Sign test

The coefficient of the independent variable represents our best estimate for the change in the dependent variable given a one-unit change in the independent variable.





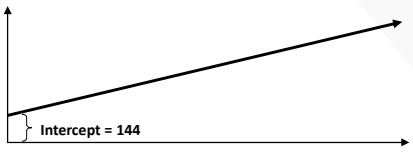
If the sign of the resulting coefficient does not match the anticipated change in the dependent variable

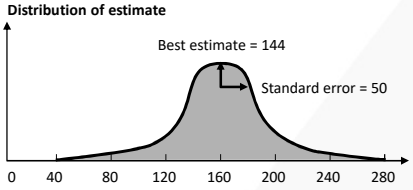
- Data may be corrupt (or incomplete) preventing the true relationship from appearing
- True relationship between variables may not be as strong as initially thought
- Counter-intuitive relationship might exist between variables

ANALYTIXLABS

Interpreting the coefficient

Similarly, the intercept represents our best estimate for the value of the dependent variable when the value of the independent variable is zero.





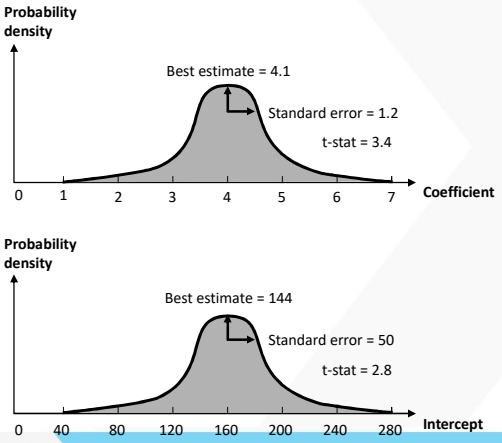
If the sign of the intercept does not match your expectation, data may be corrupt or incomplete

In some cases, it is appropriate to force the regression to have an intercept of 0, if, for instance, no meaningful value exists if the independent variable is 0

ANALYTIXLABS

T-Statistics

If the regression has passed the sign test, the single most important indicator of how strong the data supports an underlying linear relationship between the dependent and independent variables is the t-statistic.

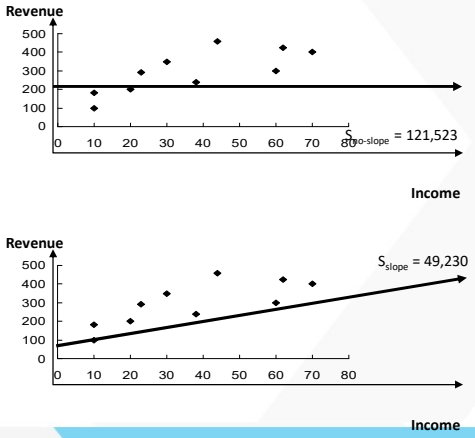


In general, a t-statistic of magnitude equal or greater than 2 suggests a statistically significant relationship between the 2 variables

ANALYTIXLABS

Interpreting R^2-Statistic

If we are comfortable with the sign and magnitude of the coefficient and intercept, and our t-statistic is sufficiently large to suggest a statistically significant relationship, then we can look at the R2-statistic.



The R^2-statistic is the percent reduction in the sum of squared residuals from using our best fit sloped line vs. a horizontal line

$$R^2 = \frac{S_{no\text{-}slope} - S_{slope}}{S_{no\text{-}slope}}$$

$$R^2 = \frac{121{,}523 - 49{,}230}{121{,}523}$$

$R^2 = 0.59$

If the independent variable does not drive (or is not correlated) with the dependent variable in any way, we would expect no consistent change in "y" with consistently changing "x." This is true when the slope is zero or $S_{slope} = S_{no\text{-}slope}$ which makes $R^2 = 0$

ANALYTIXLABS

Multiple Regression

Multiple regression allows you to determine the estimated effect of multiple independent variables on the dependent variables.

Dependent variable: Y

Independent variables:
$X_1, X_2, X_3, \ldots, X_n$

Relationship:
$Y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \ldots + a_nx_n$

Multiple regression programs will calculate the value of all the coefficients (a_0 to a_n) and give the measures of variability for each coefficient (i.e., R^2 and t-statistic)

Tests for multiple regressions

- Sign test – check signs of coefficients for hypothesized change in dependent variable
- T-statistic – check t-stat for each coefficient to establish if t>2 (for a "good fit")
- R^2, adjusted R^2 – R^2 values increase with the number of variables; therefore check adjusted R^2 value to establish a good fit (adjusted R^2 close to 1)

ANALYTIXLABS

Multiple Regression

If you can dream up multiple independent variables or "drivers" of a dependent variable, you may want to use multiple regression.

Independent variable	Dependent variables	Slopes	Intercept
y	x_1	a_1	b
	x_2	a_2	
	⋮	⋮	
	x_i	a_i	

$$y = a_1x_1 + a_2x_2 \ldots + a_ix_i + b$$
$$= b + \sum_i a_1x_1$$

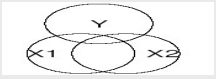
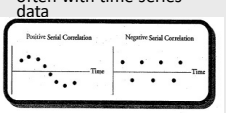
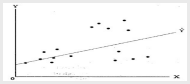
Multiple regression notes

- Having more independent variables always makes the fit better – **even** if it is not a statistically significant improvement. So:

1. Do the sign check for **all** slopes and the intercept
2. Check the t-stats (should be >2) for **all** slopes and the intercept
3. Use the adjusted R^2 which takes into account the false improvement due to multiple variables

ANALYTIXLABS

Multiple regression – 4 primary issues

	Multicollinearity	Serial correlation/ Autocorrelation	Heteroscedasticity	Outlier
What is it?	• High correlation among two or more of the independent variable 	• Residual terms are correlated with one another. It occurs most often with time series data 	• Variance of the residual term increases as the value of the independent variable 	• If some values are markedly different from the majority of the values 
Effect	• Distorts the standard error of coefficient. This will lead to greater probability of incorrectly concluding that a variable is not statistically significant (Type II error)	• Coefficient standard error too large or too small leading to erroneous t-statistic	• Standard errors will be different for different sets of independent variable	• Prediction line gets pulled-up /down in presence of outlier(s) and R-sq dips
Detection	• R-square is high, F test is statistically significant but t-tests indicate that none of the individual coefficients is significantly different than zero, VIF and CI is very high	• Scatter plot of residuals or run Durbin Watson statistic	• Examine scatter plot of residuals or run Breusch-Pagan test	• Examine from scatter plot or do an univariate analysis and look at 5,10,90,95,98,99,100 percentiles to detect outlier or check from box-plot
Correction	• Run correlation matrix and drop one of the correlated variable	• Adjust coefficient standard error using Hansen method (SAS/SPSS). This will help in correct hypothesis testing of the regression coefficient	• Calculate robust standard errors (also called White-corrected standard errors) to recalculate t-statistics	• Either drop the values or cap it by the closest observation/ replace by mean

ANALYTIXLABS

Steps in Regression Model building

1. Converting business problem into statistical problem - Identifying type of problem
2. Define hypothetical relationship (Defining Y & X variables)
3. Collect the data from across sources
4. Aggregation-getting data at same level (depends on type of problem)
5. Data Audit Report - Meta data level - table level - individual variable level
6. Data preparation
    a. Exclusions - Based business rules
    b. Data type conversions
    c. Outliers
    d. Fill rate – Missing's
    e. Derived variable creation - New variable creation - Binning of variables
    f. dummy variable creation
7. Data preparation (based on technique)
    Check the Assumptions (Y- Normal, Y & X linear)
    Transformations
    Multi-collinierity
8. Split the data into training & testing data sets(70:30)
9. Build the model on training
10. Interpreting the model - by checking few set of metrics
11. Validate the model using testing data
    1. Re-run the model
    2. Scoring the model
    3. K-Fold validation(cross Validation)
12. Preparing the final reports to share the results
13. Identify the limitations of Model
14. Converting statistical solution into Business Solution – Implementation

ANALYTIXLABS

Development of the model

Identify	Explanatory and Response variables
Decide on type of model	Here the type of model is OLS
Variable Selection	Forward Backward Stepwise
Check Multicollinearity	VIF Condition index Variance proportions
Run model	OLS
Diagnostics	For OLS
Model unsatisfactory ?	Try transformations Log, sqrt, Inverse, Box-Cox etc.

ANALYTIXLABS

Diagnostics for OLS Model

Is the model satisfactory ?

- ✓ R2 = proportion of variation in the response variable explained by the model
  - check R2 >50%
- ✓ Plots of Standardized Residual (= (Actual – Predicted)/SD)
  - vs predicted values
  - vs X variables
  - check if there is no pattern
  - check for homoscedasticity
- ✓ Significance of parameter estimates
  - check if p-value<0.01
- ✓ Stability of parameter estimates:
  - Take a random subsample from the development sample
  - Obtain a new set of parameter estimates from the sub sample
  - Check if the parameter estimates got from development sample and the subsample differ by less than 3 standard deviations
- ✓ Rank ordering:
  - -order data in descending order of predicted values
  - -Break into 10 group
  - -check if average of actual is in the same order as average predicted

ANALYTIXLABS

Validation

On the validation sample

- Stability of parameter estimates:
  - -Obtain a new set of parameter estimates from the validation sample
  - -check if the new parameter estimates differ from that got from development sample by less than 3 standard deviations
- Compare Predicted vs Actual values

ANALYTIXLABS

Regression-Best practices

1. Check for the collinearity (by finding correlation between all the variables and keeping only 1 of the variables which is highly correlated)
2. Transform data as applicable – e.g., income should be transformed by taking log of that
3. Do not run regression on categorical variables, recode them into dummy variables
4. Check the directionality of the variables
5. Following methods should be used under different situations

- **Enter Method :** To get the coefficient of each and every variable in the regression
- **Back ward method :** When the model is exploratory and we start with all the variables and then remove the insignificant ones
- **Forward Method:** Sequentially add variables one at a time based on the strength of their squared semi-partial correlations (or simple bivariate correlation in the case of the first variable to be entered into the equation)
- **Step wise method :** A combination of forward and backward at each step one can be entered (on basis of greatest improvement in R2 but one also may be removed if the change (reduction) in R2 is not significant (In the Bordens and Abbott text it sounds like they use this term to mean Forward regression)

ANALYTIXLABS

Logistic Regression

ANALYTIXLABS

Example: Brand Preference for Orange Juice

- ✓We would like to predict what customers prefer to buy: Citrus Hill or Minute Maid orange juice?
- ✓The Y (Purchase) variable is categorical: 0 or 1
- ✓The X (LoyalCH) variable is a numerical value (between 0 and 1) which specifies the how much the customers are loyal to the Citrus Hill (CH) orange juice
- ✓Can we use Linear Regression when Y is categorical?

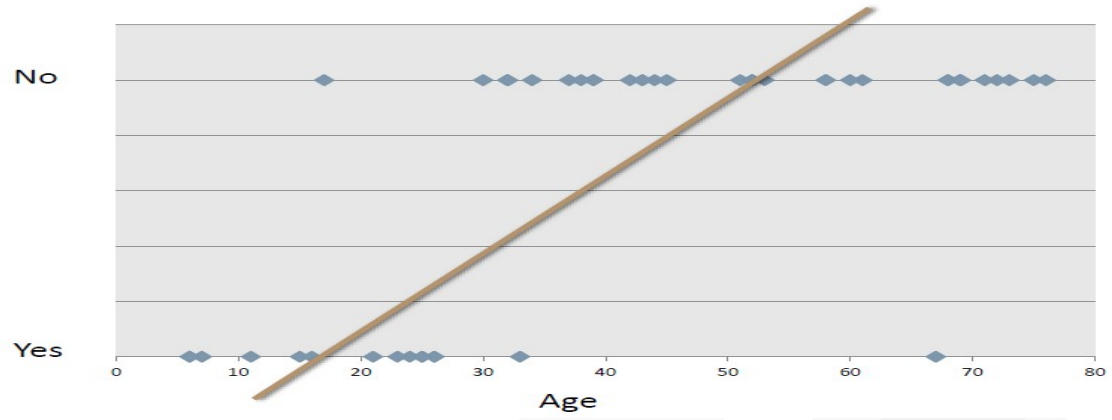
ANALYTIXLABS

Example: Credit Card Default Data

- ✓We would like to be able to predict customers that are likely to default
- ✓Possible X variables are:
  - ✓Annual Income
  - ✓Monthly credit card balance
- ✓The Y variable (Default) is categorical: Yes or No
- ✓How do we check the relationship between Y and X?

ANALYTIXLABS

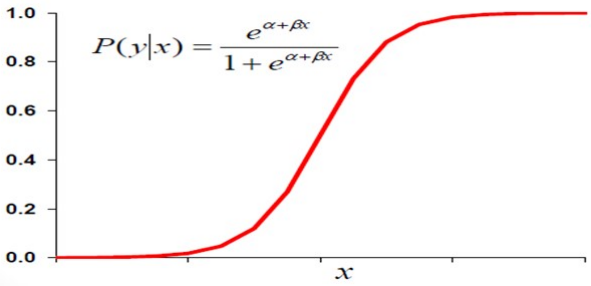
Why Not Linear?



ANALYTIXLABS

Logistic Regression

- ✓We want a model that predicts probabilities between 0 and 1, that is, S-shaped.
- ✓There are lots of S-shaped curves. We use the logistic model:
- ✓Probability = 1/[1+exp(B0+B1x) or log[p/(1-p)] =B0+B1x
- ✓The function on left, log[p/1-p], is called the logistic function



ANALYTIXLABS

Logistic Regression

- ✓Logistic regression models the logit of the outcome
  - =Natural logarithm of the odds of the outcome
  - =ln(probability of the outcome(p)/probability of not having the outcome(1-p)

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_i x_i$$

- ✓B= log odds ratio associated with predictors
- ✓Exp(B) = Odds Ratio.
- ✓The betas themselves are log-odds ratios. Negative values indicate a negative relationship between the probability of "success" and the independent variable; positive values indicate a positive relationship
- ✓Increase in log-odds for a one unit increase in x with all the other x's constant

ANALYTIXLABS

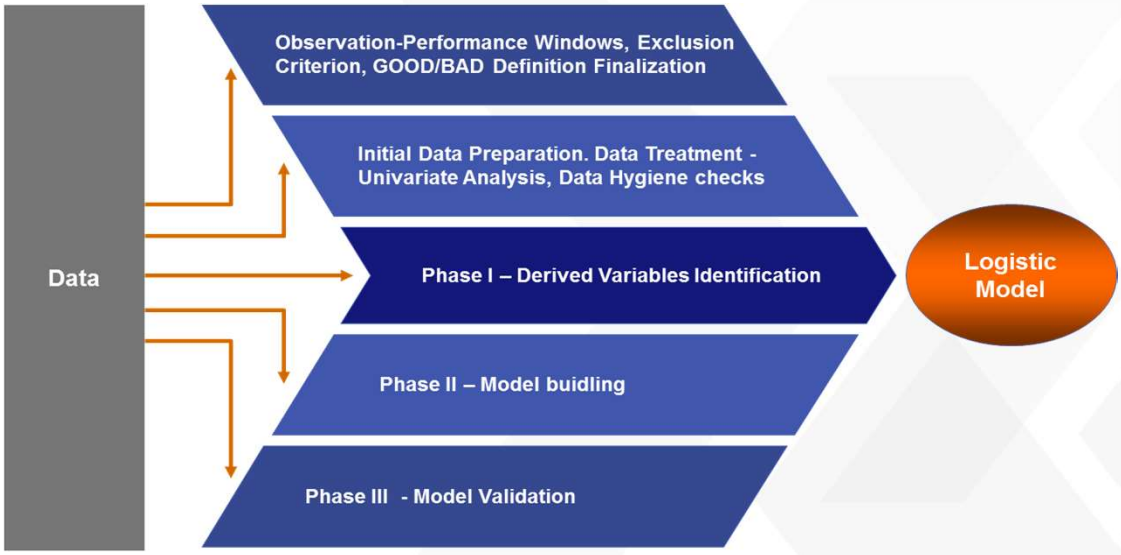
Logistic Regression

Model equation	$P_i = \text{Prob}(Y_i=0) = \frac{e^{Li}}{(1+e^{Li})}$ Where, $L_i = a+b_1X_{1i}+b_2X_{2i}+\ldots+b_pX_{pi}$
Assumption	Y_i and Y_j independent for all $i \neq j$
Parameters to be Estimated	$a, b_1, b_2, \ldots, b_p$
Method of Estimation	Maximum Likelihood

- ✓Maximum Likelihood Estimator:
  - ✓Starts with arbitrary values of the regression coefficients and constructs an initial model for predicting the observed data.
  - ✓Then evaluates errors in such prediction and changes the regression coefficients so as make the likelihood of the observed data greater under the new model
  - ✓Repeats until the model, converges, meaning the differences between the newest model and the previous model are trivial.
- ✓The idea is that you "find report as statistics" the parameters that most likely to have produced your data

ANALYTIXLABS

Modeling Methodology – Logistic Model Development



ANALYTIXLABS

Target variable(GOOD/BAD) Definition Finalization

Obs-Perf Windows

Observation window is where the Xs (independent vars) come from. Performance window is where the Y (dependent var) comes from.

Exclusion Criterion

Observations from the population that need to be excluded. Necessary to weed out data bias and to ensure model utility.

e.g. – Inactive accounts at observation point to be excluded, Accounts charged off within performance window to be excluded.

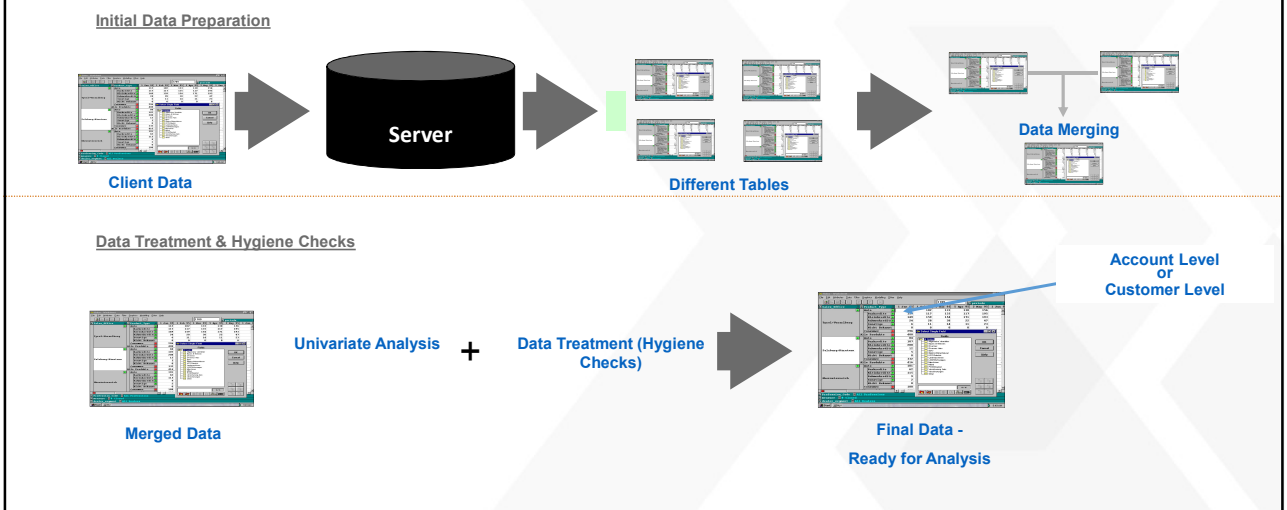
GOOD/BAD Definition

Definition of Y. How do we define a Good/Bad?

e.g. – Bad definition of 90+ ever + major derog (that includes Bankrupt, Charge offs, Repossessions, but excluding Fraud) within the Performance window. Goods are defined as anything other than a Bad.

ANALYTIXLABS

Initial Data Preparation. Data Treatment -Univariate Analysis, Data Hygiene checks



ANALYTIXLABS

Phase I – Derived Variable Identification

Raw variables could of few types – Demographic, Product Related, Behavioral, etc.

From the Raw variables (populated in the dataset) – New variables are *Derived*.

Why Derived Variables ?

- ✓ **New business relevant variables could be created by certain combinations of raw variables. *E.g. Utilization is a derived variables that is created from balance & credit limit.***
- ✓ **In certain cases aggregation variables make more sense rather than stand-alone ones. *E.g. Average payments in last 3 months, Maximum delinquency level in last 6 months…***
- ✓ **New variables creation ensures that we capture all the nuances of data.**

ANALYTIXLABS

Phase II(a) – Fine classing

- ✓ Fine classing is a process that allows us to determine which characteristics are worthy of consideration in the development of the model.
- ✓ Each characteristic is investigated to determine the underlying good/bad trends in the data at attribute level for discrete data and in small bands for continuous data.
- ✓ This process brings out the information values of the variables telling us ability of the variable to separate the goods and bads.

Log Odds (Weight of Evidence):

Log of Odds represents the proportion of Goods vis-à-vis proportion of Bads in a particular attribute. ***Weight of Evidence = ln(g/b)***

Information Value:

Information Value (IV) is a measurement of how well the characteristic can differentiate between 'good' & 'bad' and whether that characteristic should be considered for modeling.

ANALYTIXLABS

Phase II(a) – Fine classing (contd…)

Information Value:

Let g and b denote the proportion of goods and the proportion of bads for a given attribute. The following descriptive statistics are used to describe the Information Value (IV) of a particular attribute.

Information Value = [(g – b) ln(g/b)]

IV <0.03 Not Predictive **– do not consider for modeling**
IV 0.03 – 0.1 Predictive **– consider for modeling**
IV >0.1 Very Predictive **– use in modeling**

ANALYTIXLABS

Phase II(a) - Fine classing output

TABLE-TOTAL SAMPLE

acct_age	TOTAL ACCTS	ROW % TOTAL	NO. GOODS	ROW % GOODS	NO. BADS	ROW % BADS	ODDS	LOG(LN) ODDS	MARG. INFO VALUE	ROW CHI-SQUARE
Total	17204	100.00	15255	100.00	1949	100.00	1.00	0.00	0.00	0

TABLE-MARGINAL CLASSINGS

acct_age	TOTAL ACCTS	ROW % TOTAL	NO. GOODS	ROW % GOODS	NO. BADS	ROW % BADS	ODDS	LOG(LN) ODDS	MARG. INFO VALUE	ROW CHI-SQUARE
2 - 9	2065	12.00	1686	11.05	379	19.45	0.57	-0.56	0.05	101.442
10 - 16	1934	11.24	1662	10.89	272	13.96	0.78	-0.25	0.01	14.405
17 - 23	1786	10.38	1537	10.08	249	12.78	0.79	-0.24	0.01	12.139
24 - 34	1989	11.56	1743	11.43	246	12.62	0.91	-0.09	0.00	2.139
35 - 46	1919	11.15	1697	11.12	222	11.39	0.98	-0.02	0.00	0.110
47 - 66	1729	10.05	1599	10.48	130	6.67	1.57	0.45	0.02	24.985
67 - 86	1768	10.28	1675	10.98	93	4.77	2.30	0.83	0.05	64.817
87 - 121	1758	10.22	1627	10.67	131	6.72	1.59	0.46	0.02	26.307
122 - 177	1744	10.14	1581	10.36	163	8.36	1.24	0.22	0.00	6.823
178 - 422	512	2.98	448	2.94	64	3.28	0.90	-0.11	0.00	0.699

INFORMATION VALUE = 0.153

TOTAL CHI-SQUARE VALUE = 253.866 WITH 9 DF
Acct_age IS SIGNIFICANT AT THE 0 % LEVEL

ANALYTIXLABS

Phase II(b) - Coarseclassing

- ✓Coarseclassing is the grouping together of attributes of characteristics with similar performance (log odds) in the fineclassing output into coarser groups.
- ✓This allows statistically valid groupings to be modeled and allows for fluctuations within characteristics to be smoothed out. These coarse groupings are called 'dummy variables'.
- ✓In continuous variables dummies can be used to smooth a trend within a variable that deviates from the trend.

Important DOs

- Try to make classes with around 5% of the population. Classes with less than 5% might not be a true picture of the data distribution and might lead to model instability.
- Business inputs from the SMEs in the markets are essential for coarseclassing process as fluctuations in variables can be better explained and classes make business sense.

ANALYTIXLABS

Phase II(b): Dummy creation & correlation

Dummy Creation

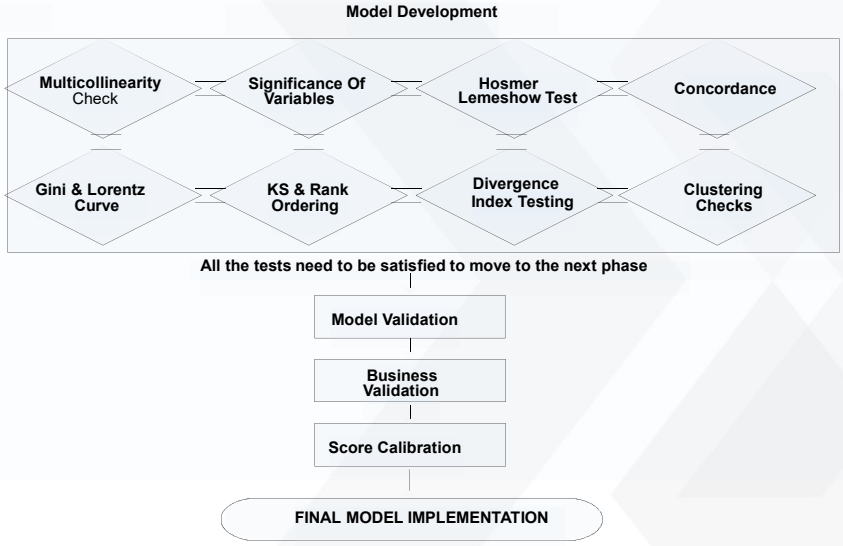
- ✓Fineclassing & Coarseclassing procedure helps in identifying the dummies to be created.
- ✓Dummying is the process of assigning a binary outcome to each group of attributes in each predictive characteristic.

Dummy Correlation Check

- ✓Once dummies are created – we need to run the correlation check on these dummies.
- ✓This is done to take care of any significant multi-collinearity effects that may exist among the dummies.
- ✓Correlation coefficient cut-off for dummy correlation is set at 0.5

ANALYTIXLABS

Phase II(c): Model Building



ANALYTIXLABS

Multicollinearity

What is Multicollinearity ?

Multicollinearity is a phenomenon when there is a linear relationship between a set of variables.

Why is Multicollinearity a problem ?

Multicollinearity affects the parameter estimates making them unreliable.

How to detect Multicollinearity ?

Variance Inflation Factor (VIF) = $1/(1- R^2)$

How to remove Multicollinearity ?

- **Look into Variance proportions table for the row with highest CI**
- **Identify variables with highest factor loadings in the row**
- **Drop the variable which is least significant**

VIF>1.75 => Multicollinearity

ANALYTIXLABS

Variable Significance

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6010	0.1423	17.8279	<.0001
d1_cons_cd_grt_1	1	1.0016	0.1326	57.0378	<.0001
d3_max_cdlevel	1	-1.0768	0.2338	21.2164	<.0001
d1_Payment_method	1	1.6529	0.1449	130.1012	<.0001
d3_OTB_jun04	1	0.6993	0.1176	35.3416	<.0001
d2_crlimit_may04	1	0.3627	0.1156	9.8523	0.0017
d2_avg_pay_bal	1	0.4720	0.1084	18.9700	<.0001
d2_max_payment	1	0.2424	0.1110	4.7691	0.0290
d4_age	1	0.4141	0.1094	14.3331	0.0002

Chi – Square value for each explanatory variable – the chi-square value indicates the level of significance, i.e – the impact of independent (explanatory) variable on the dependent variable.

The p-value cut-off should be decided in discussion with the business. Ideally the p-value<0.0001. However in case of smaller population size p-value could be <0.05 or p-value<0.1.

ANALYTIXLABS

Hosmer Lemeshow

Null Hypothesis: The expected values from the model = The observed values from the population

Alternative Hypothesis: The expected values from the model not equal to The observed values from the population

- ✓ Hosmer Lemeshow Goodness of Fit test involves dividing the data into approximately 10 groups of roughly equal size based on the percentiles of the estimated probabilities.
- ✓ The discrepancies between the observed and expected number of observations in these groups are summarized by the Pearson chi-square statistic, which is then compared to chi-square distribution with *t* degrees of freedom, where *t* is the number of groups minus 2.

Partition for the Hosmer and Lemeshow Test

Group	Total	Good = 1 Observed	Good = 1 Expected	Good = 0 Observed	Good = 0 Expected
1	924	756	753.27	168	170.73
2	1002	918	920.21	84	81.79
3	1058	997	1002.64	61	55.36
4	981	947	945.00	34	36.00
5	884	859	860.25	25	23.75
6	923	905	904.36	18	18.64
7	931	921	919.35	10	11.65
8	786	778	779.30	8	6.70
9	734	731	729.17	3	4.83
10	953	950	948.44	3	4.56

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
2.6543	8	0.9541

For a robust model – we need to *accept* the null hypothesis. Hence, *Higher the p-value better the model fit*.

ANALYTIXLABS

Concordance

Association of Predicted Probabilities and Observed Responses

Percent Concordant	79.0
Percent Discordant	19.1
Percent Tied	1.9
Pairs	3627468

- ✓Concordance is used to assess how well scorecards are separating the good and bad accounts in the development sample.
- ✓The higher is the concordance, the larger is the separation of scores between good and bad accounts.
- ✓The concordance ratio is a non-negative number, which theoretically may lie between 0 and 1.

Concordance Determination:

Among all pairs formed from 0 & 1 observations from the dependent variable, the % of pairs where the probability assigned to an observation with value 1 for the dependent variable is greater than that assigned to an observation with value 0.

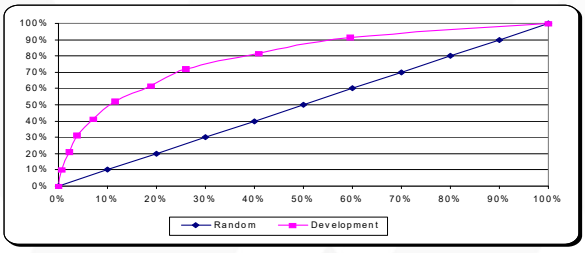
Percentage of concordant pairs should be *at least greater than 60*.

ANALYTIXLABS

Lorenz Curve, Gini, KS

Lorenz curve indicates the lift provided by the model over random selection.

Gini coefficient represents the area covered under the Lorenz curve. A good model would have a Gini coefficient between 0.2 - 0.35



Lorenz Curve

Kolmogorov-Smirnoff (KS) statistic is defined as the absolute difference of cumulative % of Goods and cumulative % of Bads.

KS statistic value *should not be less than 20*. Higher the KS – better is the model.

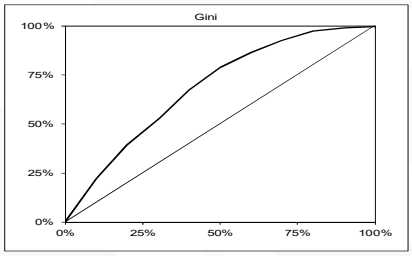
ANALYTIXLABS

Lorenz Curve, Gini, KS

Template for calculating Gini and KS coefficient for measuring discriminatory power of a rating

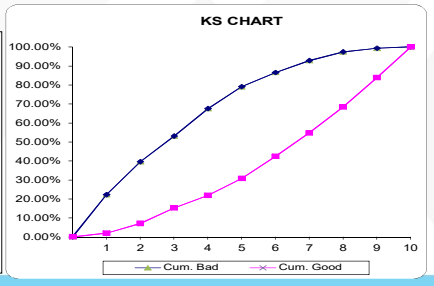
GINI-coefficient:	0.62
KS	0.48

Rating class	# of customers Good	Bad	Total	Bad rate	Accumulated pop. Bad	Good	Total	GINI area	KS
					0.00%	0.00%	0.00%		
1	12	92	104	88.46%	22.17%	1.93%	10.02%	1.11%	20.24%
2	32	72	104	69.23%	39.52%	7.06%	20.04%	3.09%	32.46%
3	51	56	107	52.34%	53.01%	15.25%	30.35%	4.77%	37.76%
4	41	60	101	59.41%	67.47%	21.83%	40.08%	5.86%	45.64%
5	55	48	103	46.60%	79.04%	30.66%	50.00%	7.27%	48.38%
6	73	31	104	29.81%	86.51%	42.38%	60.02%	8.29%	44.13%
7	77	26	103	25.24%	92.77%	54.74%	69.94%	8.89%	38.04%
8	85	19	104	18.27%	97.35%	68.38%	79.96%	9.52%	28.97%
9	96	8	104	7.69%	99.28%	83.79%	89.98%	9.85%	15.49%
10	101	3	104	2.88%	100.00%	100.00%	100.00%	9.98%	0.00%
Total	623	415	1,038	39.98%				68.65%	48.38%



Important: Rating Classes need to go from worst to best; in the example "1" is the worst rating class

In statistics, the Kolmogorov–Smirnov test (K–S test) is a nonparametric test for the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test). The two-sample KS test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.



ANALYTIXLABS

Rank Ordering

Rank Ordering is a test to validate whether the model is able to differentiate the Goods from the Bads across the population breakup.

- ✓ **The population is divided into the deciles in the descending order of predicted values (Good/Bad as the case might be).**
- ✓ **A model that rank orders, predicts the highest number of Goods in the first decile and then goes progressively down.**

Models have to rank order completely across development as well as Validation samples.

decile	Bad	Good
1	3	915
2	6	912
3	7	910
4	13	905
5	19	898
6	30	888
7	30	888
8	61	856
9	78	840
10	167	750
Total	414	8762

ranking	sat_rank
SATISFACTORY	all

ANALYTIXLABS

Divergence Index Test

Good	_FREQ_	ave	variance		Ho: Bad Score => Good Score Null Hypothesis is Rejected	p- value
	41338	752.67	4070.44			
0	856	654.55	10578.1225	DI	T - Statistic	
1	40482	754.75	3725.8816	**1.4038**	**-28.398**	**<0.0001**

Divergence Index is an indicator of how well the means of the goods and bads are differentiated.

Null Hypothesis: The means of Good accounts / population = The means of Bad accounts / population

Alternative Hypothesis: The means of Good accounts / population is not equal to the means of Bad accounts / population

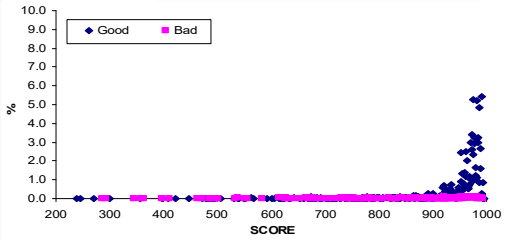
For a robust model – we need to *reject* the null hypothesis. Hence, *lower the p-value better the model*.

ANALYTIXLABS

Clustering check

The concept behind Clustering check is that a good model should be sensitive enough to differentiate between 2 Good/Bad accounts.

i.e the model should be able to identify differences between seemingly same type of accounts/sample observations and assign them different scores.



A good model should not have significant clustering of the population at any particular score and the population must be well scattered across.

Ideally the clustering should be as low as possible. A thumb-rule would be to contain the clustering so that it is *within 5-6%*.

ANALYTIXLABS

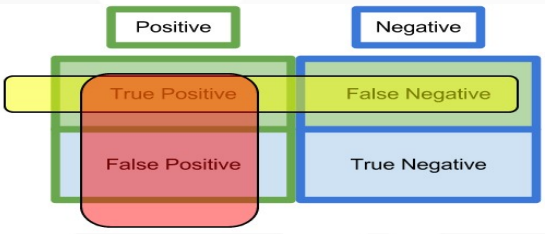
Other Metrics

Coefficient's signs & stability	**Coefficients signs** must match in the models run on both the samples. **Stability** (significance and parameter estimates should be within 95% Confidence limits of parameter estimates) in the models run on both the samples.
Divergence Index	$D = \frac{\bar{x}_A - \bar{x}_B}{s}$ is a commonly employed measure of the separation achieved by a model. It is related to a t-distribution (multiply by (G+B) ½) if the two population variances are equal. This measure how well the means of the respondents and non-respondents are differentiated. A t statistic > |6| shows a high level of differentiation.
Somers' D	It is used to determine the strength and direction of relation between pairs of variables. Its values range from -1.0 (all pairs disagree) to 1.0 (all pairs agree). It is defined as $(n_c-n_d)/t$ where n_c is the number of pairs that are concordant, n_d the number of pairs that are discordant, and t is the number of total number of pairs with different responses.
Gamma	The Goodman-Kruskal Gamma method does not penalize for ties on either variable. Its values range from -1.0 (no association) to 1.0 (perfect association). Because it does not penalize for ties, its value will generally be greater than the values for Somer's D.
Kendall's Tau-a	It is a modification of Somer's D that takes into the account the difference between the number of possible paired observations and the number of paired observations with a different response. It is defined to be the ratio of the difference between the number of concordant pairs and the number of discordant pairs to the number of possible pairs $(2(n_c-n_d)/(N(N-1))$. Usually Tau-a is much smaller than Somer's D since there would be many paired observations with the same response.

Confusion Metrics

CONFUSION MATRIX

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative



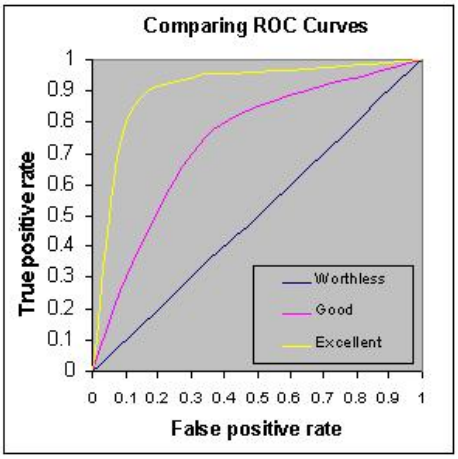
sensitivity = recall = tp / (tp + fn)
specificity = tn / (tn + fp)
precision = tp / (tp + fp)

Sensitivity/recall – how good a test is at detecting the positives.
Specificity – how good a test is at avoiding false alarms.
Precision – how many of the positively classified were relevant.

Receiver Operating Characteristic Curve: Plot of TPR(Sensitivity) vs FPR(1- Specificity)

ROC Curve

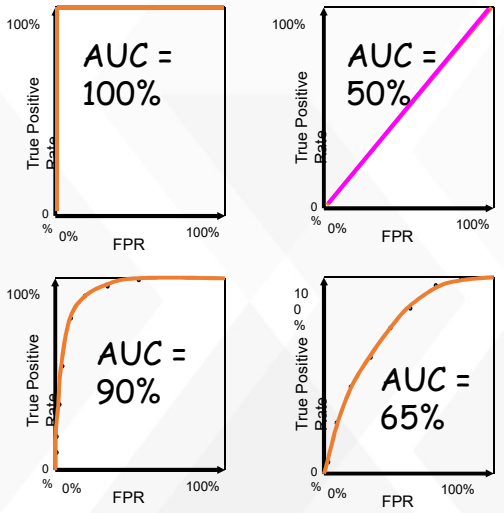
- *ROC* = *Receiver Operating Characteristic*
- Started in electronic signal detection theory (1940s - 1950s)
- Plot of TPR(Sensitivity) vs FPR(1-Specificity)
- Can be used in machine learning applications to assess classifiers



ANALYTIXLABS

ROC Curve - AUC

- *Overall measure* of model performance
- In classification,
- AUC = Concordance+0.5*Ties

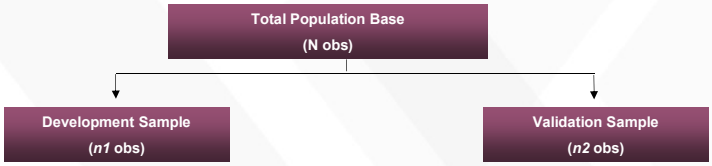


ANALYTIXLABS

Phase III: Model-Validation

Validation could be done in 2 ways:

- ✓ **Validation Re-run**
- ✓ **Scoring the Validation sample**



Validation Re-run

- Rerun the model on the validation sample.
- Check the chi-sq values and level of significances and p-values for each explanatory variable.
- The p-values should not change significantly from the development sample to the validation sample.
- Check the signs of the parameter estimates. They should not change from development sample to the validation sample.
- Check rank ordering. Both Development and validation samples should rank order.

Validation sample scoring

- Score the validation sample using the parameter estimates obtained from the scorecard developed on the development sample.
- Check rank ordering. Both development and validation samples should rank order.

ANALYTIXLABS

Model Evaluation

- Model validity refers to the stability and reasonableness of the logistic regression coefficients.
- The plausibility and usability of the fitted logistic regression function.
- The ability to generalize inferences drawn from the analysis.
- For model validation following statistical measures can be compared between the development and validation sample.
  - Coefficient sign's & Stability
  - Concordance/Somer's D
  - Decile Analysis/Rank Ordering
  - ROC Curve/Gini Coefficient
  - Kolmogorov–Smirnov test (K–S test)
  - Classification Matrix

ANALYTIXLABS

Steps to check stability of model between training and validation

Check	Test	Results
• Predictive power	Overall Gini	The overall Gini measure is 70%*, which is very good.
• Consistency in rank-ordering	Visual assessment of bar charts	Model rank-orders response consistently.
• Variable validity	Statistical significance of variables Plausibility of coefficient signs	All model variables are significant. Direction of all variables' effects plausible.
• Model stability	Out-of-sample stability of coefficients Multicollinearity tests (correlation and VIF)	New data would yield quasi identical model. No dangerous levels of correlated variables found.
• Model calibration	Correlation with actual bad rate	High correlation between predicted and actual bad rate

ANALYTIXLABS

Appendix

ANALYTIXLABS

Rare events description and example

Rare Events:

- Certain group or event happens very rarely and so its incidence in the data is very sparse and effort needs to be made to make sure they are well represented in the sample.
- Use **stratified sampling** method for rare events.
- Keep all (or most) of the observations for the rare events but sample the non-rare events more heavily.
- Calculation adjustment needs to be done to determine actual ratio between the rare and other events .
- Examples- Fraud, email campaigns , churn etc

ANALYTIXLABS

Sampling Techniques when there is Low Response Rate (rare events)

Biased Sampling	Methodology
Biased sampling is a non-random sampling procedure that incorporates a systematic bias/error in sample selection. It generates a statistical sample of a population where some members of the population are more likely to be included than others. This would imply that some members are underrepresented or overrepresented relative to others in the population.	1.Create two datasets, one having events and other having non-events data. 2. All the events are used in modeling. 3. From non-events base data, pick up that many observations randomly such that event rate based on all events and random selection of non-events data be equal to desired event rate. Post the model is developed, the bias is adjusted using a correction factor (ratio of log of odds of sample to log of odds of population).

Assigning Weights	Methodology

ML estimator for logistic regression gives equal weight to type 1 and type 2 error
If only a few percent of the sample are response (mirroring the population), estimator focuses on predicting "non-response"
However, biggest economic impact (loss) is caused by response accounts
By changing weight to 50:50, model tries to better predict "response", and economic performance of model can be improved
	1.Calculate the response percentage in the overall population
	2. Then compute(decide) the weights such that the sample would have the response and non-response in the same proportions
	3. Create weight variable as follows
	multiplier=(**100**-response-percent)/response-percent;
	if response_flag =1 then weight.= multiplier;
	else weight.=1;

ANALYTIXLABS

Bias adjustment when we use Biased Sampling

If the event rate is as low as 0.05% then the event rate is increased to about 5% (desired event rate).
How would we increase: by keeping all the events data and part of non-events is randomly picked from non-events such that new event rate is about 5%.

Whenever a bias sample is used in the model development, it's suggested to carry out 'Bootstrapping' and 'Jacknifing' at the time of model validation. These two practices would help to check if there is any bias in the parameter estimation.

p_s is the sample response rate (e.g., 5%); p_p is the actual population response rate (historical or, better, predicted future).

Logit score

$$Y = c + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

Logistic score (Biased)

$$P = \frac{1}{1 + e^{-Y}} = \frac{e^{Y}}{e^{Y} + 1}$$

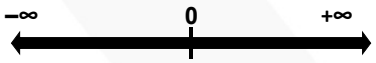
Calibrated Score

$$P^* = \frac{P * P_p * (1-P_s)}{P * P_p * (1-P_s) + (1-P) * (1-P_p) * P_s}$$

Calibration adjustment when we use Biased Sampling

Logit score

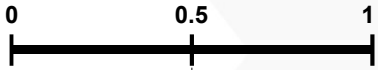
$$Y = c + \beta_1 X_1 + \beta_2 X_2 + \ldots$$



- Logit score is unbounded
- For the average case (X value is sample mean for each X), Y is typically 0*
- If "response" was coded as 1, then higher Y is higher probability of response

Logistic score

$$P = \frac{1}{1 + e^{-Y}} = \frac{e^{Y}}{e^{Y} + 1}$$



- Logistic score is bounded between 0 and 1
- P = 0.5 typically corresponds to the sample response rate*
- If "response" was coded as 1, then higher P is higher probability of response

Calibrated Score

$$P^* = \frac{\frac{p_p}{p_s} P}{\frac{p_p}{p_s} P + \frac{1 - p_p}{1 - p_s} (1 - P)}$$

- P* is bounded between 0 and 1 (i.e., 0% and 100%)
- However, numerical value now corresponds to the estimated Probabilities

* Assuming model was built on a sample with 50/50 split between response and non-response cases
** p_s is the sample response rate (e.g., 50%); p_p is the actual population response rate (historical or, better, predicted future).

Boot Strapping & Jackknifing – Validation

Boot Strapping

Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution. One standard choice for an approximating distribution is the empirical distribution of the observed data. In the case where a set of observations can be assumed to be from an independent and identically distributed population, this can be implemented by constructing a number of resamples of the observed dataset (and of equal size to the observed dataset), each of which is obtained by random sampling with replacement from the original dataset.

Jackknifing

Jackknifing, which is similar to bootstrapping, is used in statistical inferencing to estimate the bias and standard error in a statistic, when a random sample of observations is used to calculate it. The basic idea behind the jackknife estimator lies in systematically recomputing the statistic estimate leaving out one observation at a time from the sample set. From this new set of "observations" for the statistic an estimate for the bias can be calculated and an estimate for the variance of the statistic.

SOURCE: Wikipedia Images

ANALYTIXLABS

ANALYTIXLABS

Contact us

Visit us on: http://www.analytixlabs.in/

For course registration, please visit: http://www.analytixlabs.co.in/course-registration/

For more information, please contact us: http://www.analytixlabs.co.in/contact-us/
Or email: info@analytixlabs.co.in
Call us we would love to speak with you: (+91) 88021-73069

Join us on:
Twitter - http://twitter.com/#!/AnalytixLabs
Facebook - http://www.facebook.com/analytixlabs
LinkedIn - http://www.linkedin.com/in/analytixlabs
Blog - http://www.analytixlabs.co.in/category/blog/

ANALYTIXLABS