insAnalytics

Consulting. Training. Research & Development.

Insights.. Through Analytics…

Data Mining

Discriminant Analysis

© insAnalytics

Key Objectives

After successful completion of the topic, participants will be able to:

- Develop awareness of the utility of **Discriminant Analysis** as a classification technique and **key difference with other** classification techniques
- Articulate the **key concepts** of Linear Discriminant Analysis, such as **assumptions**, **"discriminant functions"** , **"discriminant scores"**, deciding the cut-off score of discriminant score etc.
- **Develop** linear discriminant model and **classify** observations using the model
  - For response variable with two levels
  - For response variable with more than two levels

To learn & understand the subject matter better, participants need to be aware of the following areas:

- Descriptive Statistics
- Statistical Estimation
- Testing of Hypotheses
- Analysis of Variance (ANOVA)
- Data Mining – CRISP DM Methodology
- Logistic Regression
- Decision Tree
- Cluster Analysis

© insAnalytics

Discriminant Analysis

- Introduction
- Assumptions and Terminologies
- Discriminant Model Development Methodology
- Building a Discriminant Analysis Model
- Classification Using Discriminant Function
- Classification with More than Two Categories

© insAnalytics

Discriminant Analysis

© insAnalytics

Discriminant Analysis

Introduction

- A Typical Classification Problem
- Graphical Illustration
- What is Discriminant Analysis
- Comparison between Different Classification Tools
- How Does It Work

Assumptions and Term

Discriminant Model D

Building a Discriminan

Classification Using Discriminant Function

Classification with More than Two Categories

© insAnalytics

Product Affinity Analysis– A Hypothetical Problem

Vikrant is managing the sales team of *Reliable Insurance*, an non-life insurance provider.

Reliable Insurance has 3 kinds of primary insurance products:

- ✓ Motor Insurance
- ✓ Fire Insurance
- ✓ Travel Insurance

For long, Vikrant was looking for a tool or **rule engine that helps his team to know which customer is inclined to buy which product**, so that their sales effort is more smooth and conversion rate is high.

© insAnalytics

Product Affinity Analysis– A Typical Classification Problem

Vikrant approached to Pratibha, the marketing analyst, with the problem.

I want to assign each of my leads into segments based on their potential interest on our products.

Vikrant, Head Sales

Pratibha, Marketing Analyst

This means you want to **classify your customer into some segments, that you already know!!**

© insAnalytics

Product Affinity Analysis– A Typical Classification Problem

But, according to your sales experience, which could be the **influencing factors**? What do you think?

Vikrant, Head Sales

Pratibha, Marketing Analyst

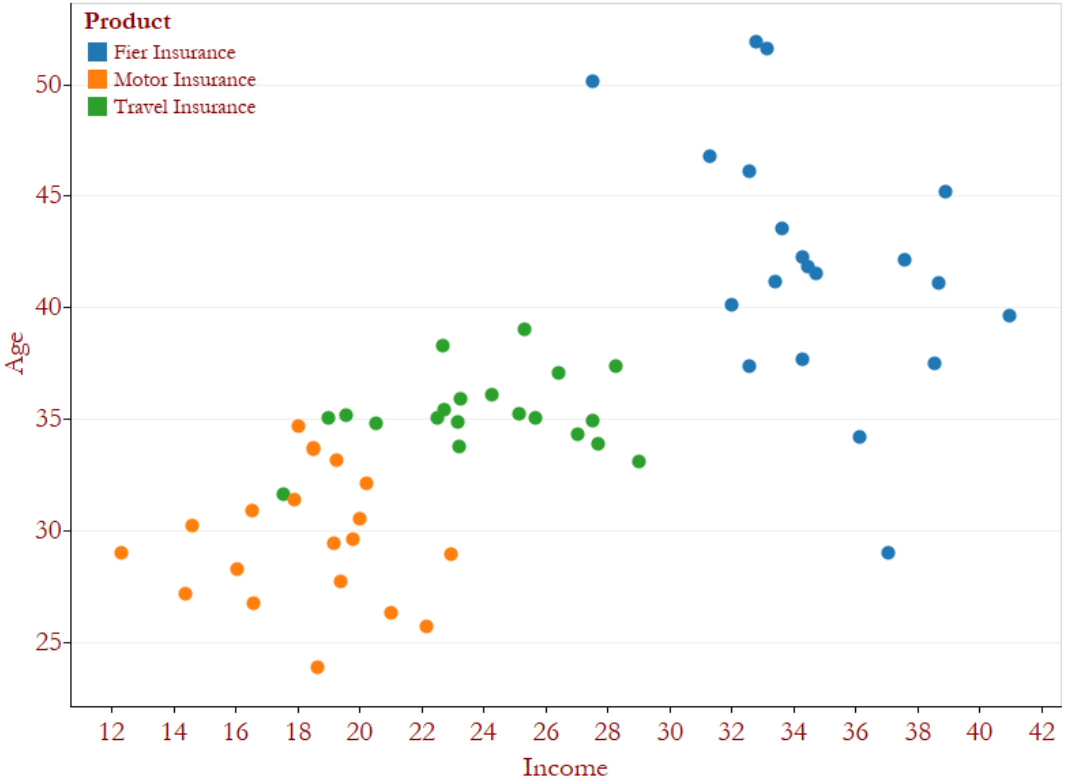
I think "Age "and the "Yearly Income "can be two major influencing factors

© insAnalytics

Product Affinity Analysis– Graphical Illustration

Pratibha took sample of 20 converted leads (i.e. first time buyer) from each product and plotted their age and yearly income (in Lac rupees).

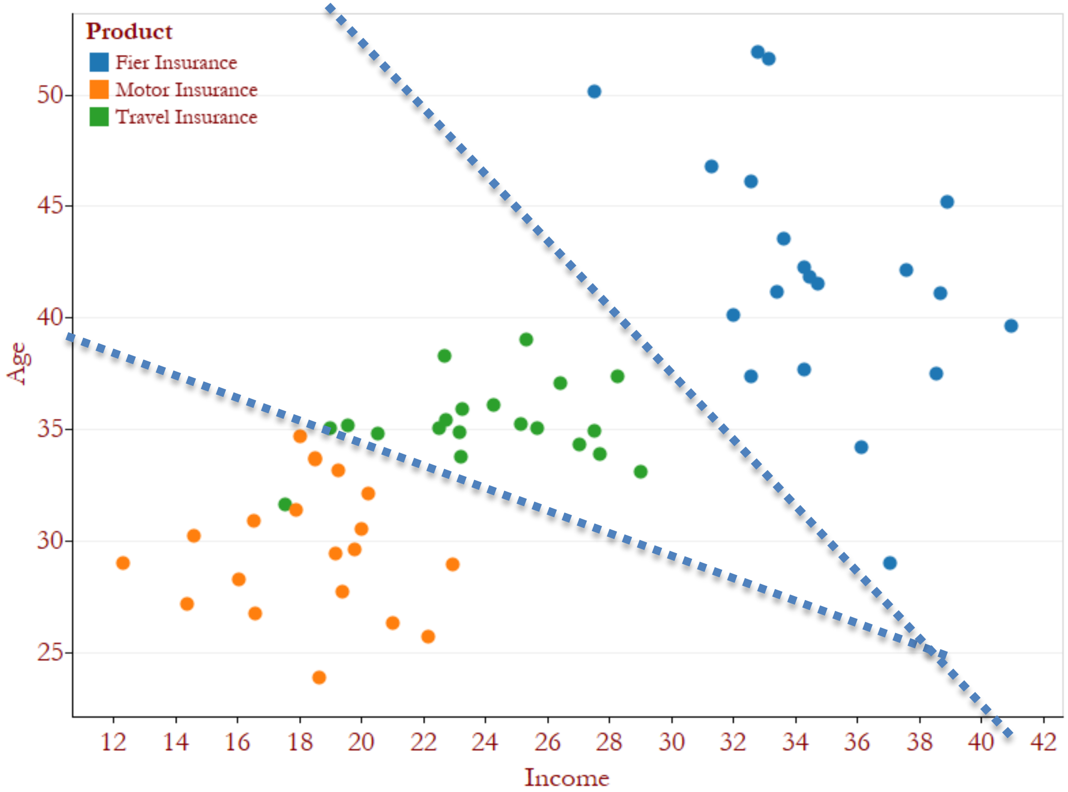
She got the following plot..



© insAnalytics

Product Affinity Analysis– Graphical Illustration

All she needs to do is to define some **boundaries with respect to "Age" and "Income"** variables, to define the segments.

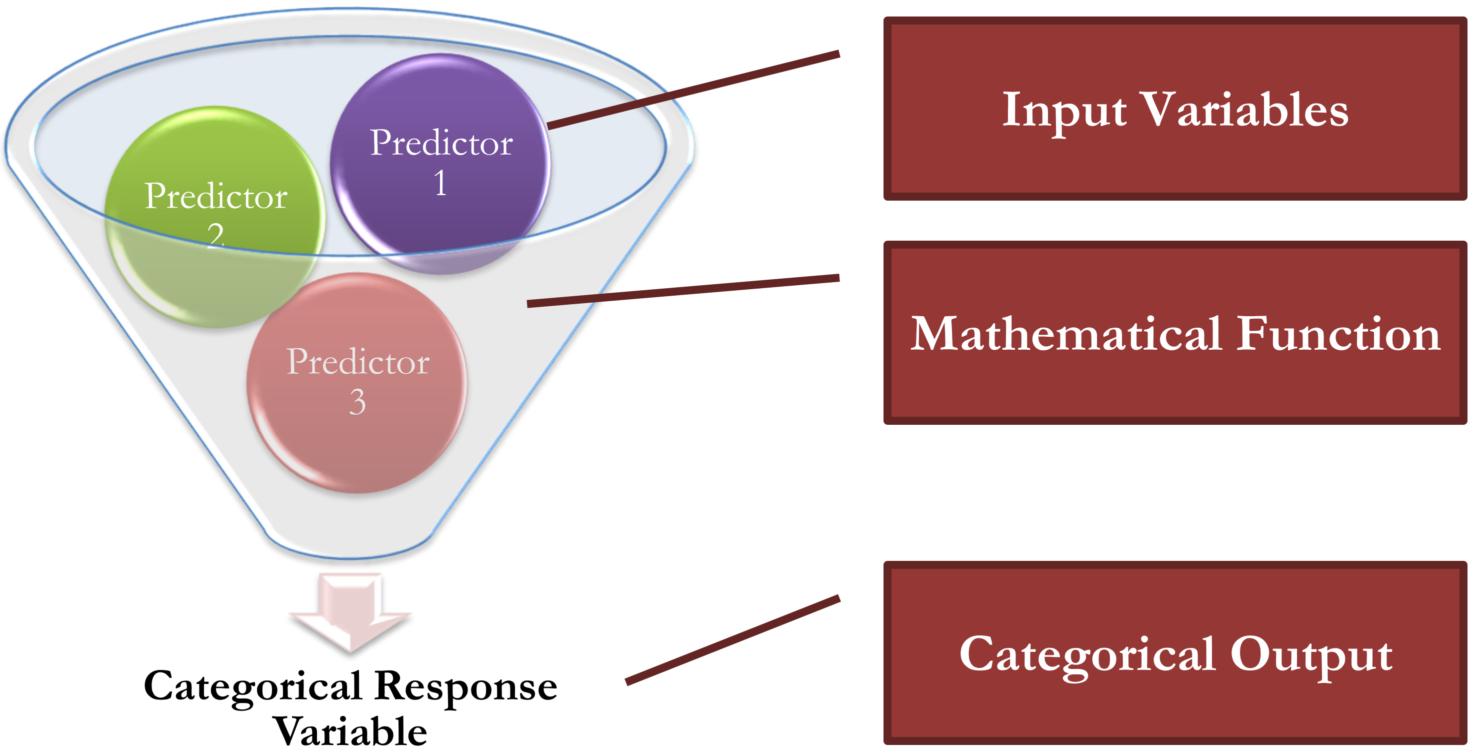


Discriminant Analysis helps to create such boundaries

© insAnalytics

Discriminant Analysis – Definition

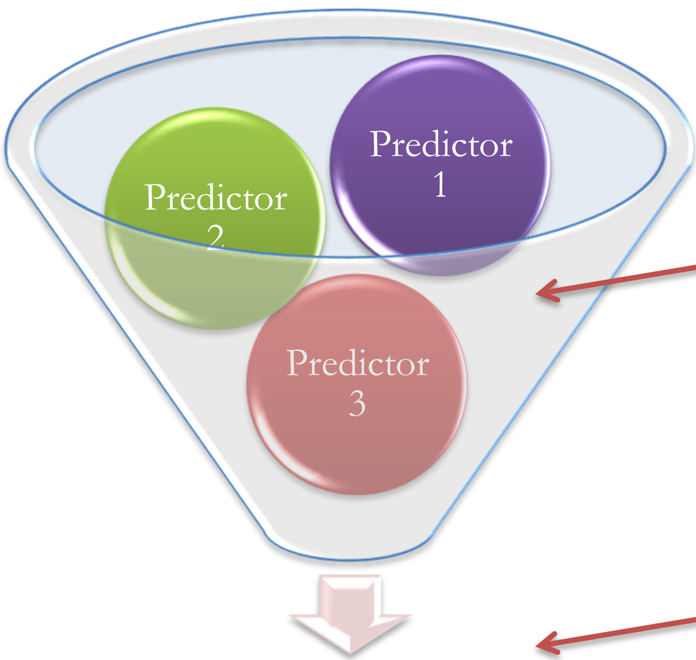
Discriminant analysis (DA) is a method to find one or more **mathematical functions** of influencing or **independent variables** that characterize(s) or **separates two or more classes of objects** or events, i.e. a categorical response variable



© insAnalytics

Discriminant Analysis – Definition

Discriminant analysis (DA) is a method to find one or more mathematical functions of influencing or independent variables that characterize(s) or separates two or more classes of objects or events, i.e. a categorical response variable



Categorical Response Variable

If the function is a linear combination, it is called **Linear Discriminant Analysis (LDA)**

If the number of categories are more that two, it is called **Multivariate Discriminant Analysis (MDA)**

© insAnalytics

Comparison Between Different Classification Tools

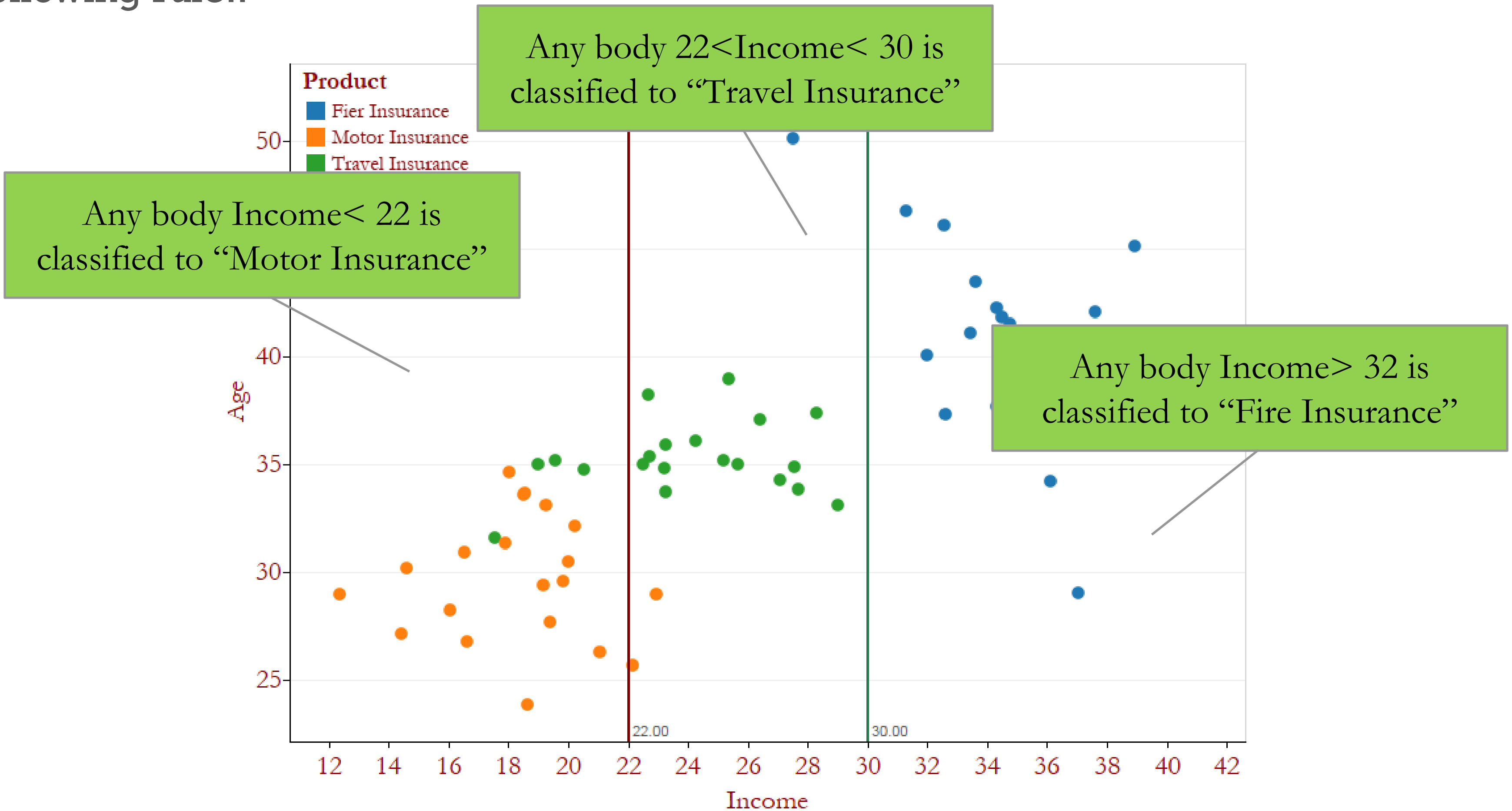
There are a number techniques available to solve a classification problems. Let's look at how they are different

	Logistic Regression	Decision Tree	Discriminant Analysis
Binary Output Categories	✓	✓	✓
More than two Output Categories	✗	✓	✓
Mathematical function as classifier	✓	✗	✓
Distributional assumptions	✓	✗	✓
Categorical Predictors	✗	✓	✗
Assigns probability to each observation	✓	✗	✗

© insAnalytics

How Does It Work – An Illustration

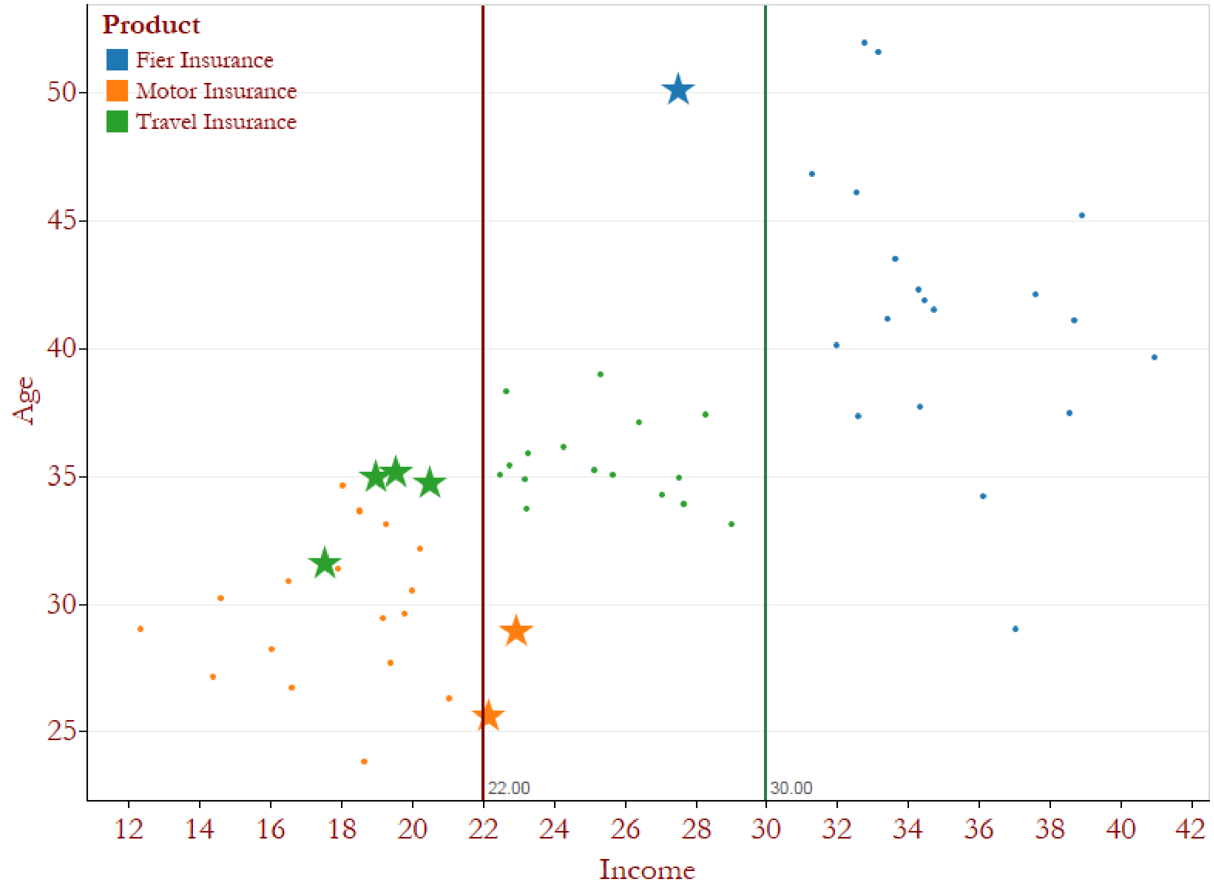
Suppose Pratibha decided to have "Income" variable as the classifier and uses the following rule!!



© insAnalytics

How Does It Work – An Illustration

But, 7 observations are **misclassified**!!! Hence, she needs to change the rule!



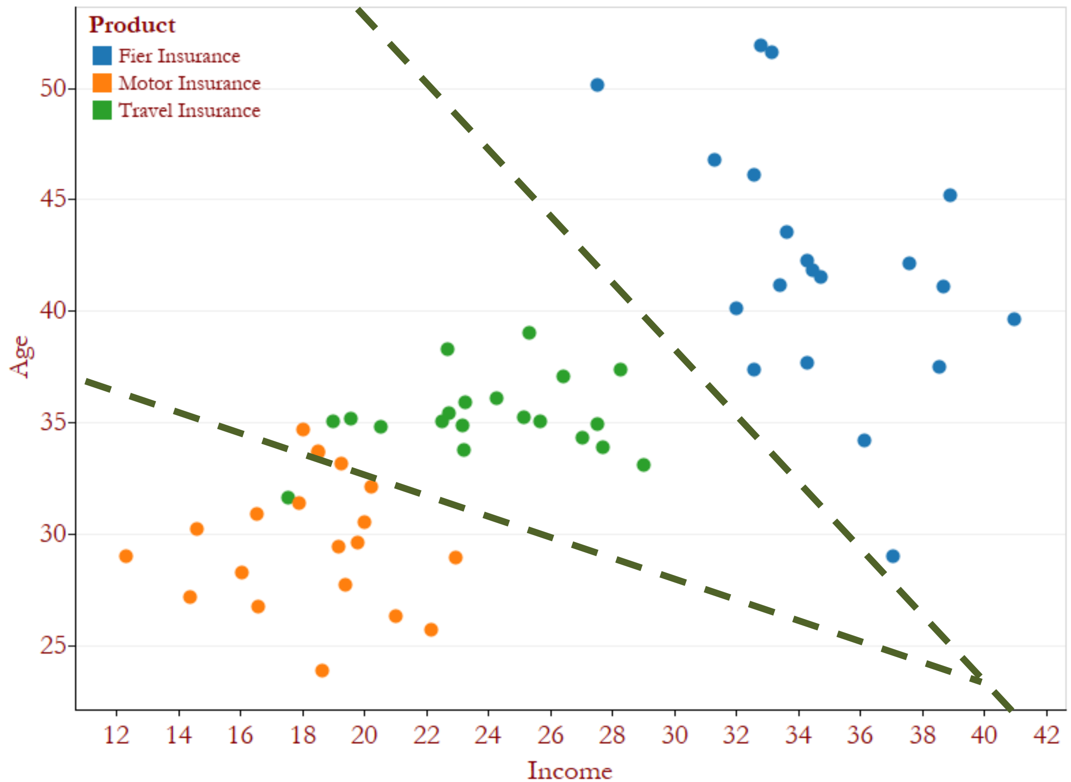
© insAnalytics

How Does It Work – An Illustration

Finally Pratibha find the following two lines as a classifier:

Function1: $2*Age + Income = 85$

Function 2: $2*Age + 3*Income = 160$



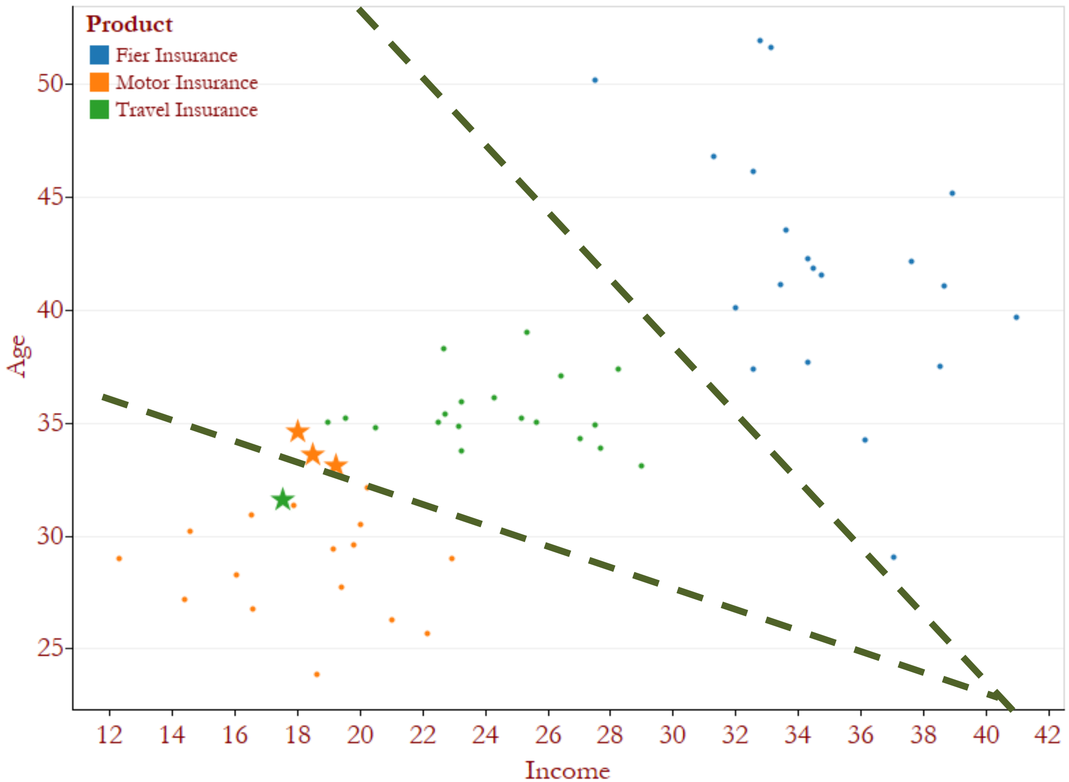
© insAnalytics

How Does It Work – An Illustration

Function1: 2*Age + Income =85

Function 2: 2*Age + 3*Income =160

With this, misclassification has reduced to 4!!! Hence, this is a better the rule!!



As the classification functions are linear combinations of the variables, the analysis is called a **Linear Discriminant Analysis**

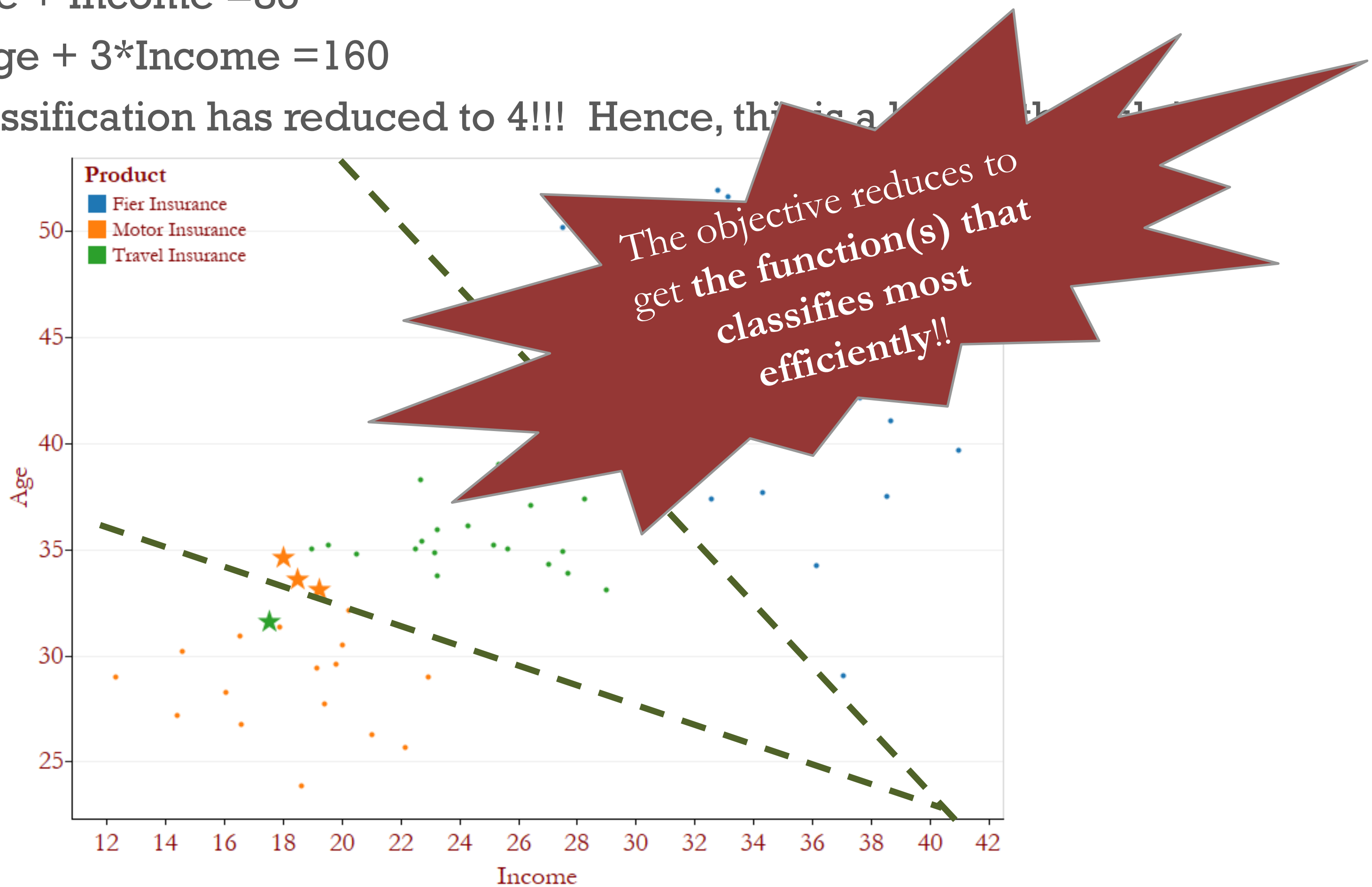
© insAnalytics

How Does It Work – An Illustration

Function1: 2*Age + Income =85

Function 2: 2*Age + 3*Income =160

With this, misclassification has reduced to 4!!! Hence, thi



As the classification functions are linear combinations of the variables, the analysis is called a **Linear Discriminant Analysis**

© insAnalytics

Discriminant Analysis

- Introduction ✓
- **Assumptions and Terminologies**
- Discriminant Model Development Methodology
- Building a Discriminant Analysis Model
- Classification Using Discriminant Function
- Classification with More than Two Categories

© insAnalytics

Discriminant Analysis

- Introduction ✓
- **Assumptions and Terminologies**
  - Assumptions
  - Discriminant Functions
  - Discriminant Score
  - Cut-off Score
- Discriminant Model Development Methodology
- Building a Discriminant Function
- Classification Using Discriminant Function
- Classification with More than Two Categories

© insAnalytics

Assumptions

There are few assumptions on the variables, both input and output, for DA:

- ✓ The input or **independent variables** has to be **numeric**
  - – Hence, the variables like gender, educational qualification, customer type can not be used as independent variables
- ✓ The output or **target variable** has to be **categorical** (with two or more). Preferably, it should be nominal variable. But, ordinal variables also can be used
  - – The target variables are like good/bad, preference of an individual, species of a biological specimen, etc.
- ✓ **Group membership** of an individual **should be known** before hand, i.e. which category of response it falls in.
  - – If we are trying to classify insurance claims as fraud or not, we should know which observations in the development data are fraud and which are not

© insAnalytics

Assumptions

There are few assumptions on the variables, both input and output, for DA:

- ✓ The input variables follow **normal distribution**
- ✓ Input variables must not be collinear, i.e. **multicollinearity should not exist**
- ✓ Input variables should have **same variance across categories**
- ✓ **At least two observations** for each category of response
- ✓ Number of observations should be **at least two more than number of predictors**

© insAnalytics

Product Affinity Analysis – Defining the Data

In the previous example of Reliable Insurance, let us try to formulate the problem mathematically.

The response variable is "Product that will Convert", which has 3 levels, viz. Motor Insurance, Travel Insurance, and Fire Insurance.

This is a categorical variable with 3 levels. ☑

The possible variables may be

- Age of the individual
- Yearly Income
- Household size
- Number of Independent, etc.

Suppose there are 'p' such variable, all of which are numeric. ☑

Lets call these variables as $X_1, X_2, \ldots, X_p$

© insAnalytics

Product Affinity Analysis –Assumptions

As per the assumptions for DA, we will assume,

We know in the dataset which lead has taken which product. Hence the group membership is known ☑

Input variables, i.e. X_1, X_2, …., X_p, follow normal distribution ☑

X_1, X_2, …., X_p are independent (at least not perfect linear relationship) ☑

Variance-of these variables , i.e. X_1, X_2, …., X_p, are same across 3 groups, i.e. products ☑

If 'n' is the number of observations in the dataset, n> p+2 ☑

© insAnalytics

Terminologies – Discriminant Function

Hence, the objective reduces to getting a set of linear combinations (say, D_i) which will discriminate the groups.

$$D_i = \beta_{0i} + \beta_{1i} * X_1 + \beta_{2i} * X_2 + \ \ldots + \beta_{pi} * X_p$$

Each of these functions are called **discriminant function**

Hence,

Discriminant functions are the combinations of the input variables those discriminate between the target groups. Values of these functions are used to assign one observation into its group.

© insAnalytics

Terminologies – Discriminant Function

But how many discriminant functions can be there?

It can be **at most one less than the number of categories** in the target variable.

For example, in the previous example, there are 3 groups, i.e. products. Hence, there can be **at most** 2 discriminant functions.

© insAnalytics

Terminologies – Discriminant Function

But how many discriminant functions can be there?

It can be **at most one less than the number of categories** in the target variable.

For example, in the previous example, there are 3 groups, i.e. products. Hence, there can be **at most** 2 discriminant functions.

If the **two or more groups can not be discriminated** through the input variables, then number of **discriminant functions are less** that its maximum possible value.

© insAnalytics

Terminologies – Discriminant Scores

Value of the discriminant function for one individual is called **discriminant score** for the observation.

© insAnalytics

Terminologies – Discriminant Scores

Value of the discriminant function for one individual is called **discriminant score** for the observation.

Suppose following is a discriminant function

$$D_1 = 20 + Age + 2 * Income - 8 * Household\ Size$$

© insAnalytics

Terminologies – Discriminant Scores

Value of the discriminant function for one individual is called **discriminant score** for the observation.

Suppose following is a discriminant function

D_1= 20+ Age+ 2* Income - 8*Household Size

Suppose, for one individual

Age= 22 Income= 26.5 Household Size= 3

For that individual, the discriminant score will be

20+22+2*26.5- 8*3

= 74

© insAnalytics

Terminologies – Discriminant Scores

Value of the discriminant function for one individual is called **discriminant score** for the observation.

Suppose following is a discriminant function

D_1= 20+ Age+ 2* Income - 8*Household Size

Suppose, for one individual

Age= 22 Income= 26.5 Household Size= 3

For that individual, the discriminant score will be

20+22+2*26.5- 8*3

= 74

The group where the individual we will fall in **is based on the discriminant score**

© insAnalytics

Terminologies – Cut-off Scores

The discriminant score provides a basis of distributing the observations into the categories or segments defined by the target variable.

Hence, we need to decide a value of the discriminant score above which we will define the observations into one group, else we assign to other group.

This value, **based on which we discriminate the observations** into categories are called the cut-off score.

It is needless to say deciding the cut-off score is as critical as deciding the discriminant functions.

© insAnalytics

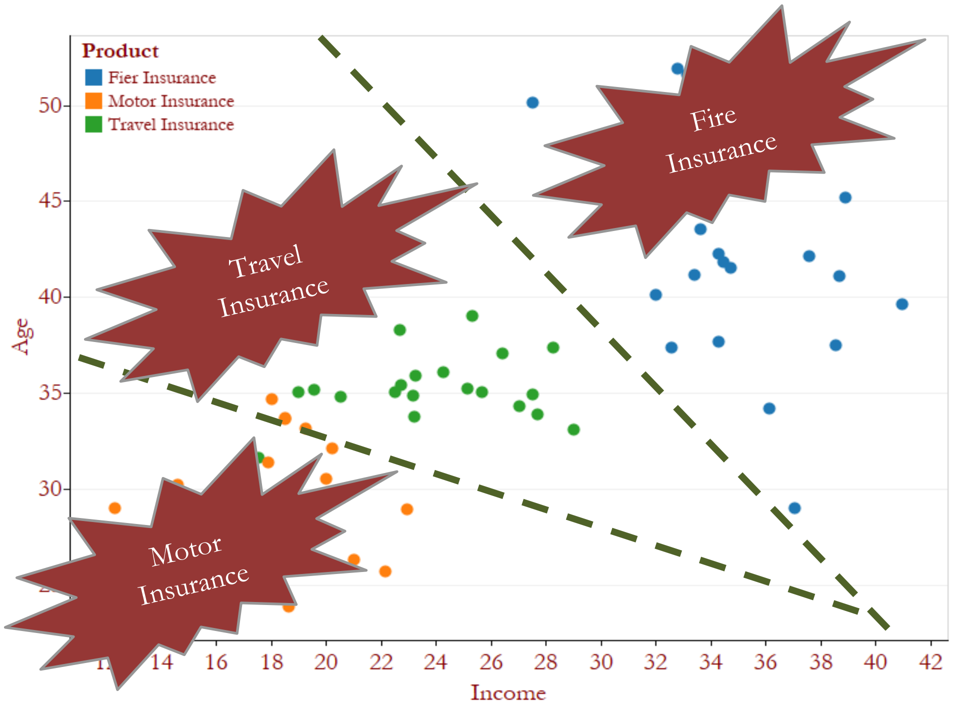
Product Affinity Analysis– Illustration

In the previous example, we had discriminant functions as

D1: 2*Age + Income

D2: 2*Age + 3*Income

The decided cut-off score for D1 was 85 and that for D2 was 160.



Hence, the rule was as below:

If D1<85 then

"Motor Insurance"

If D1>=85 and D2<160 then

"Travel Insurance"

If D2>=160 then

"Fire Insurance"

© insAnalytics

Discriminant Analysis

- Introduction ✓
- Assumptions and Terminologies ✓
- **Discriminant Model Development Methodology**
- Building a Discriminant Analysis Model
- Classification Using Discriminant Function
- Classification with More than Two Categories

© insAnalytics

Discriminant Analysis

- Introduction ✓
- Assumptions and Terminologies ✓
- **Discriminant Model Development Methodology**
- Building a Discriminant Analysis Model
- Classification Using Discriminant Function
- Classification with More than Two Categories

© insAnalytics

Development Methodology

Define
- Select the objective of the problem
- Define categories w. r. t. the target variable
- Data collection

Design
- Identify the independent variables
- Sample size considerations (if any)
- Collection of data and creation of holdout or validation sample

Explore
- Data sanity check and variable treatment
- Normality of independent variables
- Homoscedasticity check

Model
- Estimation of discriminant functions
- Interpretation of the functions

Evaluate
- Prediction accuracy
- In-time and Out of time validation

© insAnalytics

Mapping with CRISP DM Stages

Stage	Activities	CRISP DM Stage
Define	• Select the objective of the problem • Define categories w. r. t. the target variable • Data collection	Business Understanding
Design	• Identify the independent variables • Sample size considerations (if any) • Collection of data and creation of holdout or validation sample	Data Understanding and Data Preparation
Explore	• Data sanity check and variable treatment • Normality of independent variables • Homoscedasticity check	Data Preparation
Model	• Estimation of discriminant functions • Interpretation of the functions	Modeling
Evaluate	• Prediction accuracy • In-time and Out of time validation	Evaluation

© insAnalytics

Discriminant Analysis

- Introduction ✓
- Assumptions and Terminologies ✓
- Discriminant Model Development Methodology ✓
- **Building a Discriminant Analysis Model**
- Classification Using Discriminant Function
- Classification with More than Two Categories

© insAnalytics

Discriminant Analysis

© insAnalytics

e-Commerce Retailer Rating Challenge

SHOPYOURNEED is an online marketplace where retailers meet their buyers. The company decided to give a certificate to the listed retailers based on feedback from the buyers.

Currently they are doing this manually. But, they want to build a rule engine to classified the retailers as "Certified" or "Not Certified"

There are 2 categories and hence it is a **binary classification**. There will be only one discriminant function.

© insAnalytics

e-Commerce Retailer Rating Challenge– The Data

The data contains close to 50K retailers details and their final classification as "Certified" or "Not Certified".

Following features are used as possible predictors:

Average rating in last 12 months given by customer on

- Product Quality	- Speed of Delivery
- Competitive Pricing	- Complain Resolution

Average rating given by SHOPYOURNEED on

-Product Line	- Number of Complains per 100 sale

The data is provided on a normalized scale.

© insAnalytics

Understanding the Predictors – Centroid

First we need to see if these variables are at all efficient to predict response variable or not.

We would expect mean of the variables to be significantly different across categories.

We check this for the data and here is the result:

	Certified	Not Certified
Product Quality	0.6768157	0.3698688
Speed of Delivery	0.3534292	0.2847268
Competitive Pricing	0.03248880	0.0835166
Complain Resolution	0.5107170	0.4491026
Product Line	0.4799183	0.4680861
No of Complains	0.1379808	0.3540847

These means form **centroids** of the categories.

© insAnalytics

Understanding the Predictors – Centroid

In this case clearly,

✓Average "Product Line" is not different between the groups.

✓Also, "Speed of Delivery" or "Complain Resolution" is not so different.

✓Other 4 variables are significantly different

	Certified	Not Certified	
Product Quality	0.6768157	0.3698688	☑
Speed of Delivery	0.3534292	0.2847268	?
Competitive Pricing	0.03248880	0.0835166	☑
Complain Resolution	0.5107170	0.4491026	?
Product Line	0.4799183	0.4680861	☒
Number of Complains	0.1379808	0.3540847	☑

© insAnalytics

Understanding the Predictors – Between vs. Within Variation

We will also expect a good predictor will have high between group variation.

We express between group variation as percentage of total variation. We call this , **omega square** (Ω^2).

We would expect Ω^2 to be as high as possible.

As part of calculation, we can run ANOVA on each of the predictor response variable as class variable. Then we calculate Omega square as,

$$\Omega^2 = 1 - SSE/SST$$

$$= 1 - (\text{Error Sum of Squares}/\ \text{Total Sum of Squares})$$

© insAnalytics

Understanding the Predictors – Between vs. Within Variation

The result is as below

	Between SS	Within SS	Total SS	Ω^2
Product Quality	574.9870	1733.9847	2308.9717	24.9%
Speed of Delivery	28.8055	298.4678	327.2733	8.8%
Competitive Pricing	15.8908	386.6419	402.5327	3.9%
Complain Resolution	23.1684	1994.1306	2017.2990	1.1%
Product Line	0.8544	886.0598	886.9142	0.1%
Number of Complains	285.0078	4456.0508	4741.0586	6.0%

Ideally we would expect the **Omega Square to be more than 10% or 5%**

Hence, "Product Line", "Complain Resolution", "Competitive Pricing" are found to be not significant

© insAnalytics

Selecting the Predictors– Estimation Methods

As an alternative, we can select variables using any of the following two estimation methods.

Simultaneous Method

- It considers all the input variables concurrently to develop discriminant function

Stepwise Estimation

- It selects predictors one by one into discriminant function to arrive the best discriminant function

But often people **advise to avoid "Stepwise Estimation"**, just to ensure the right kind of business variables are selected.

Alternatively we **look at the centroids and omega square** for the variables.

© insAnalytics

Building a Discriminant Analysis Model – R Code

Looking at the centroids and omega square, we use the following 3 variables for linear discriminant analysis.

- ✓ **Product Quality** – Centroids are significantly different, omega square is also high
- ✓ **Competitive Pricing**– Centroids are different though omega is not very high ~4%
- ✓ **Number of Complains** – Centroids are different though omega is not very high ~6%
- ✓ **Speed of Delivery** – Though centroids are closers, omega is decently high ~9%

The R code for building the model is

```
fit <- lda(Response ~
          Product.Quality+ Speed.of.Delivery+
          Number.of.Complains+ Competetive.Pricing, data=online,
          na.action="na.omit")
```


© insAnalytics

Building a Discriminant Analysis Model – R Code

The function for Linear Discriminant Analysis

The response variable, **has to be a categorical variable**

```
fit <- lda(Response ~
         Product.Quality+ Speed.of.Delivery+
         Number.of.Complains+ Competetive.Pricing, data=online,
         na.action="na.omit")
```

The predictor variables, **has to be numeric**

Ensures list-wise deletion in case of missing observations

The data for analysis

© insAnalytics

Building a Discriminant Analysis Model – The Output

```
Prior probabilities of groups:
              Certified Not Certified
              0.1162998 0.8837002
Group means:
              Product.Quality Competetive.Pricing Speed.of.Delivery
Certified     0.6768157       0.1377911            0.3534292
Not Certified 0.3698688       0.3534553            0.2847268

              Number.of.Complains
Certified     0.1379808
Not Certified 0.3540847

Coefficients of linear discriminants:
                              LD1
Product.Quality          -5.130362
Competetive.Pricing      -0.121136
Speed.of.Delivery        -8.205428
Number.of.Complains       0.704665
```


© insAnalytics

Building a Discriminant Analysis Model – The Output

```
Prior probabilities of groups:
              Certified Not Certified
              0.1162998 0.8837002
Group means:
              Product.Quality Competetive.Pricing Speed.of.Delivery
Certified     0.6768157       0.1377911           0.3534292
Not Certified 0.3698688       0.3534553           0.2847268

              Number.of.Complains
Certified     0.1379808
Not Certified 0.3540847

Coefficients of linear discriminants:
                              LD1
Product.Quality          -5.130362
Competetive.Pricing      -0.121136
Speed.of.Delivery        -8.205428
Number.of.Complains       0.704665
```

Prior Probabilities: Original proportion in the data

© insAnalytics

Building a Discriminant Analysis Model – The Output

```
Prior probabilities of groups:
             Certified Not Certified
             0.1162998 0.8837002
```

```
Group means:
             Product.Quality Competetive.Pricing Speed.of.Delivery
Certified     0.6768157        0.1377911            0.3534292
Not Certified 0.3698688        0.3534553            0.2847268

             Number.of.Complains
Certified     0.1379808
Not Certified 0.3540847
```

Centroids

```
Coefficients of linear discriminants:
                                LD1
Product.Quality          -5.130362
Competetive.Pricing      -0.121136
Speed.of.Delivery        -8.205428
Number.of.Complains       0.704665
```


© insAnalytics

Building a Discriminant Analysis Model – The Output

```
Prior probabilities of groups:
             Certified Not Certified
             0.1162998 0.8837002
Group means:
             Product.Quality Competetive.Pricing Speed.of.Delivery
Certified     0.6768157       0.1377911           0.3534292
Not Certified 0.3698688       0.3534553           0.2847268

             Number.of.Complains
Certified     0.1379808
Not Certified 0.3540847
```

```
Coefficients of linear discriminants:
                                LD1
Product.Quality             -5.130362
Competetive.Pricing         -0.121136
Speed.of.Delivery           -8.205428
Number.of.Complains          0.704665
```

Discriminant function

© insAnalytics

Building a Discriminant Analysis Model – The Output

```
Prior probabilities of groups:
             Certified Not Certified
             0.1162998 0.8837002
Group means:
             Product.Quality Competetive.Pricing Speed.of.Delivery
Certified    0.6768157       0.1377911           0.3534292
Not Certified 0.3698688      0.3534553           0.2847268

             Number.of.Complains
Certified    0.1379808
Not Certified 0.3540847
```

```
Coefficients of linear discriminants:
                             LD1
Product.Quality         -5.130362
Competetive.Pricing     -0.121136
Speed.of.Delivery       -8.205428
Number.of.Complains      0.704665
```

Discriminant function

The coefficients are also called "**Discriminant Loadings**"

© insAnalytics

Discriminant Analysis

- Introduction ✓
- Assumptions and Terminologies ✓
- Discriminant Model Development Methodology ✓
- Building a Discriminant Analysis Model ✓
- **Classification Using Discriminant Function**
- Classification with More than Two Categories

© insAnalytics

Discriminant Analysis

- Introduction ✓
- Assumptions and Terminologies ✓
- Discriminant Model Development Methodology ✓
- Building a Discriminant Analysis Model ✓
- **Classification Using Discriminant Function**
  - Calculation of Discriminant Score
  - Classification and Cutting Score
  - Optimal Cut-off Score
  - Classification Matrix
  - Hit Ratio
- Classification with More than Two Categories

© insAnalytics

Calculation of Discriminant Score

Discriminant analysis is a classification technique and **discriminant scores form the basis of classification**.

Once discriminant function is available, we calculate discriminant score for each observation.

In the previous example, discriminant score can be calculated by:

D1= (-5.130362)*Product.Quality + (-0.121136)*Competetive.Pricing +(-8.205428)*Speed.of.Delivery + (0.704665)*Number.of.Complains

In R we can calculate the discriminant score by:

```
predicted<-predict(fit, newdata=online)$x
```

Where 'fit' is the discriminant model we got previously

© insAnalytics

Classification and Cutting Score

Now the question comes
"What should be the cut-off score?"

Let's look at how the calculated discriminant score is distributed across two categories.

Group	Min	P10	P20	P25	P50	P75	P80	P90	Max
Certified	-6.7964	-2.7004	-2.4202	-2.3199	-1.9470	-1.6466	-1.5739	-1.4050	1.1972
Not Certified	-2.9761	-1.1066	-0.7259	-0.5574	0.2312	1.0348	1.2218	1.6989	3.7572

Hence, there is a overlap region on the values of discriminant scores from two groups.

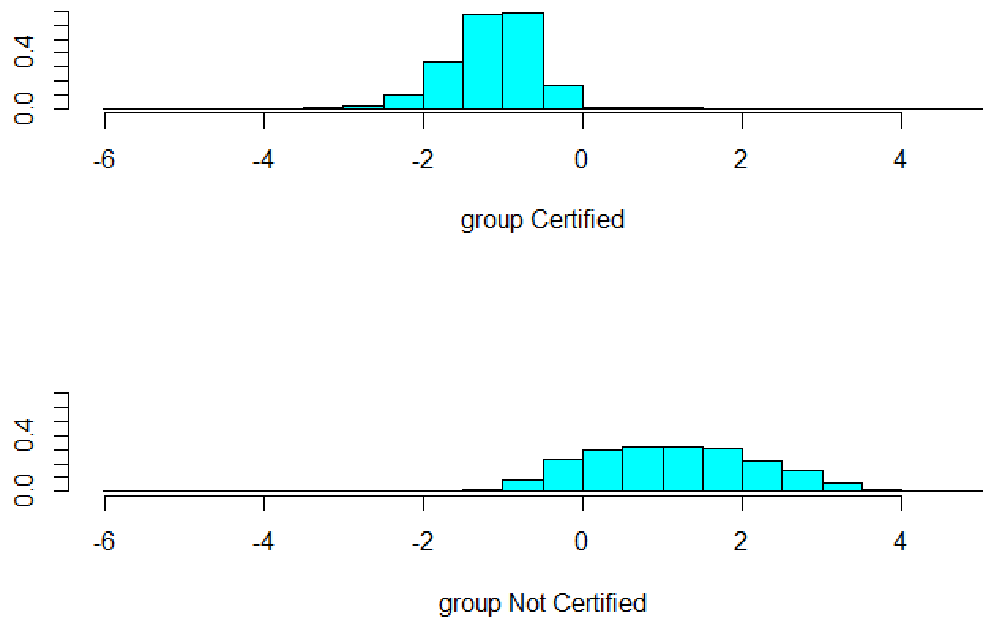
If we select 1.19 (the maximum value for "Certified") as cut off, almost 77-78% "Not Certified" will also be marked as "Certified".

So, which value should we get

© insAnalytics

Classification and Cutting Score

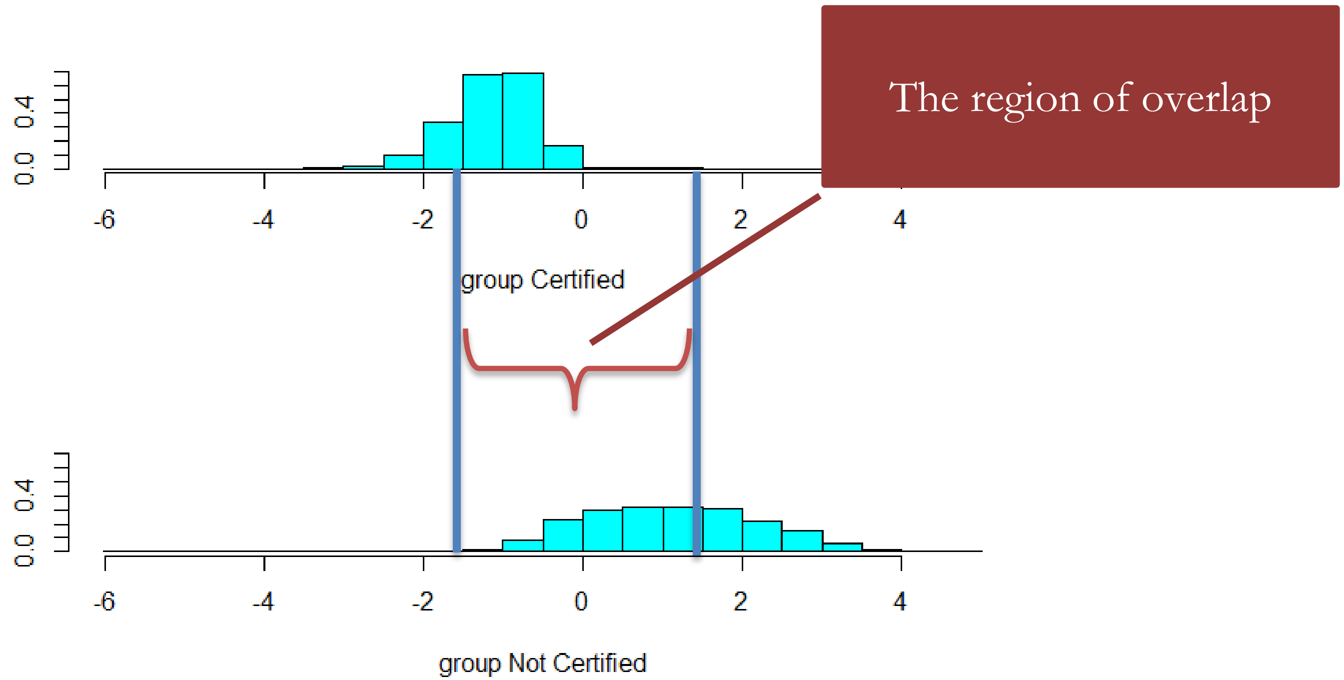
If we use the R code `plot(fit)` , we get the following plot.



© insAnalytics

Classification and Cutting Score

If we use the R code `plot(fit)` , we get the following plot.



Any point can be a cut-off point. But which one is optimal?

© insAnalytics

Optimal Cutting Score

Let us look at the distribution of scores in two categories.

We define,

Z_A= Average score for Group A, also called centroid of A
Z_B= Average score for Group B , also called centroid of B
N_A= Number of observations in Group A
N_B= Number of observations in Group B

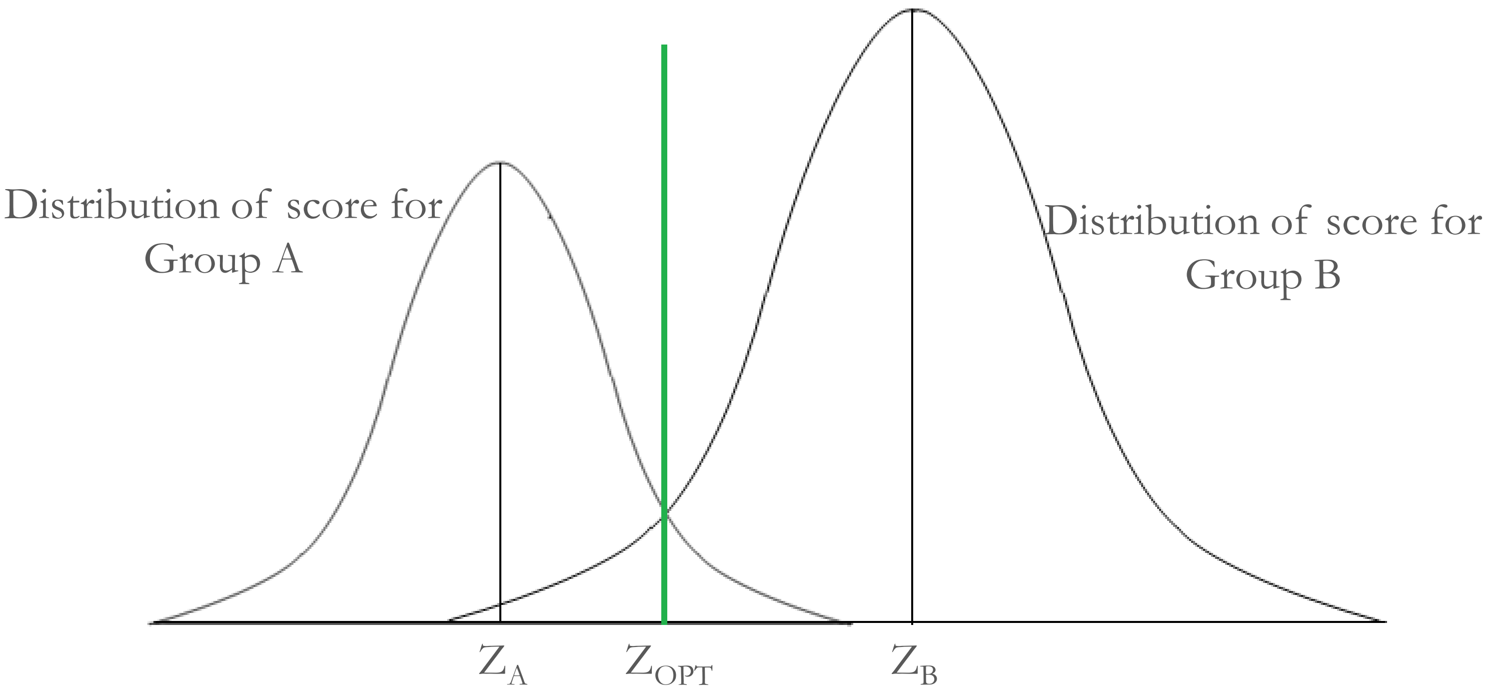
The optimal cut-off point is defined as:

$$Z_{OPT}= (Z_B*N_A+Z_A*N_B)/(N_A+N_B)$$

© insAnalytics

Optimal Cutting Score- Graphical Illustration

Let us look at the distribution of scores in two categories.



© insAnalytics

Optimal Cutting Score – Example

In the previous example,

Z_{CERT}= -2.0123 $Z_{NOT\ CERT}$= 0.2634
N_{CERT}= 6,906 $N_{NOT\ CERT}$=52,475

The optimal cut-off point is :

Z_{OPT}= -1.7476

Hence,
any **retailer with score less than -1.7476 will be classified as "Certified"**,
else they will classified as "Not Certified"

We can directly get this classification using the R code;

```
predicted_class<-predict(fit, newdata=online)$class
```


© insAnalytics

Classification Matrix– Basis of Effectiveness Measure

As in any classification problem like decision tree or logistic regression, here also the basic measure of classification efficiency is the classification matrix. Confusion matrix

We calculate the classification matrix as below:

	Classified "Certified"	Classified "Not Certified"	Total
Actual "Certified"	4,591	2,315	6,906
Actual " Not Certified"	505	51,970	52,475
Total	5,096	54,285	59,381

© insAnalytics

Classification Matrix– Basis of Effectiveness Measure

As in any classification problem like decision tree or logistic regression, here also the basic measure of classification efficiency is the classification matrix.

We calculate the classification matrix as below:

Wrong Decision

Right Decision

	Classified "Certified"	Classified "Not Certified"	Total
Actual "Certified"	4,591	2,315	6,906
Actual " Not Certified"	505	51,970	52,475
Total	5,096	54,285	[illegible]81

Wrong Decision

Right Decision

© insAnalytics

Measuring Efficiency – Hit Ratio

We have defined "**Accuracy Rate**" or "**Hit Ratio**" before (refer Logistic Regression). We calculate the same here as well.

	Classified "Certified"	Classified "Not Certified"	Total
Actual "Certified"	4,591	2,315	6,906
Actual " Not Certified"	505	51,970	52,475
Total	5,096	54,285	59,381

Hit Ratio = Correct Decision/ Total number of observations

= (4591+51970)/59381

= **95.25%**

© insAnalytics

Discriminant Analysis

- Introduction ✓
- Assumptions and Terminologies ✓
- Discriminant Model Development Methodology ✓
- Building a Discriminant Analysis Model ✓
- Classification Using Discriminant Function ✓
- **Classification with More than Two Categories**

© insAnalytics

Discriminant Analysis

© insAnalytics

e-Commerce Retailer Rating Challenge – The Problem

SHOPYOURNEED believes there are some retailers who are possible fraud and making the good-will of the marketplace at stake.

Hence, they decided to classify the retailers in to 3 categories, instead of two. The classes being,

- "Certified" – those who have proven to have performed well
- "Restricted" – those who will not be allowed list any product, as they have shown some fraudulent behavior
- "Not Certified" – those who does not fall on either of the above two

How their analysis will change now?

As, there are 3 categories there will be **at most two discriminant functions**.

© insAnalytics

Understanding the Predictors – Centroid

We use the same data as before and like before we first look at the centroids.

	Certified	Not Certified	Restricted	
Product Quality	0.6768157	0.4051427	0.2991655	☑
Speed of Delivery	0.3534292	0. 2869413	0.2802879	☑
Competitive Pricing	0.1377791	0.1872670	0.6865465	☑
Complain Resolution	0.5107170	0.4566559	0.4339626	☒
Product Line	0.4799183	0.4791408	0.4459280	☒
Number of Complains	0.1379808	0.1875742	0.6878398	☑

Clearly "Product Line" and "Complain Resolution" are not significant enough.

© insAnalytics

Understanding the Predictors – Between vs. Within Variation

Now we look at the omega square calculation

	Between SS	Within SS	Total SS	Ω^2
Product Quality	705.859	1603.073	2308.932	30.6%
Speed of Delivery	29.321	297.946	327.267	9.0%
Competitive Pricing	3188.491	1552.058	4070.549	67.3%
Complain Resolution	29.169	1988.096	2017.265	1.4%
Product Line	13.708	873.191	886.889	1.5%
Number of Complains	3201.239	1539.739	4740.978	67.5%

Here, also our conclusion remain same.

Hence, we run discriminant analysis with 4 variables, viz.
Product Quality, Speed of Delivery, Competitive Pricing, and Number of Complains

© insAnalytics

Discriminant Analysis Model – The Output

```
Prior probabilities of groups:
     Certified    Not Certified     Restricted
     0.1162998    0.5895657         0.2941345

Group means:
             Product.Quality Competetive.Pricing Speed.of.Delivery Number.of.Complains
Certified     0.6768157        0.1377911            0.3534292          0.1379808
Not Certified 0.4051427        0.1872760            0.2869413          0.1875742
Restricted    0.2991655        0.6865465            0.2802879          0.6878398

Coefficients of linear discriminants:
                      LD1          LD2
Product.Quality       2.7620415    4.85526134
Competetive.Pricing    0.2190726   0.03368713
Speed.of.Delivery     1.9978335    9.03860856
Number.of.Complains  -6.0022724    2.29167587

Proportion of trace:
     LD1          LD2
     0.8831       0.1169
```


© insAnalytics

Discriminant Analysis Model – The Output

```
Prior probabilities of groups:
     Certified   Not Certified    Restricted
     0.1162998   0.5895657        0.2941345

Group means:
             Product.Quality Competetive.Pricing Speed.of.Delivery Number.of.Complains
Certified     0.6768157         0.1377911           0.3534292           0.1379808
Not Certified 0.4051427         0.1872760           0.2869413           0.1875742
Restricted    0.2991655         0.6865465           0.2802879           0.6878398
```

```
Coefficients of linear discriminants:
                      LD1            LD2
Product.Quality       2.7620415      4.85526134
Competetive.Pricing    0.2190726     0.03368713
Speed.of.Delivery     1.9978335      9.03860856
Number.of.Complains  -6.0022724      2.29167587
```

Note, two discriminant functions are calculated!!

```
Proportion of trace:
     LD1          LD2
     0.8831       0.1169
```


© insAnalytics

Which Discriminant Function is More Powerful?

Now we may be interested to know which discriminant function is more powerful?

This can be calculated by the output "Proportion of trace", which tells **percentage of between-group variance explained by each discriminant function**.

Let's look at the output again:

```
Coefficients of linear discriminants:
                      LD1            LD2
Product.Quality       2.7620415      4.85526134
Competetive.Pricing    0.2190726     0.03368713
Speed.of.Delivery     1.9978335      9.03860856
Number.of.Complains   -6.0022724     2.29167587
```

```
Proportion of trace:
     LD1          LD2
     0.8831       0.1169
```

The first discriminant function explains 88.31% variation

© insAnalytics

When Do We Get Less Discriminant Function than Expected?

Till now we know, if there are k categories we will get at most (k-1) discriminant functions.

But, when we will have less than (k-1) discriminant functions?

Suppose a case where we have 4 categories and we have the following trace values:

```
Proportion of trace:
            LD1      LD2       LD3
            0.8728   0.1201    0.0071
```

Do we need the 3rd discriminant function? It only explains 0.71% of variation?

What does that mean?

Two categories, out of 4, will **remain indistinguishable** through the analysis using the variables under consideration

© insAnalytics

Potency Index– How the Variables Influencing the Functions

The analyst may want to see how the variables are impacting the discriminant functions. For that we calculate Potency Index

Potency Index of a variable with respect to a discriminant function is defined as

Potency Index = Squared "discriminant loading" divided by "Relative Eigen Value" for the discriminant function

Relative Eigen Value = Eigen value divided by total of all eigen values

We can get the eigen value (also called singular value) of the discriminant function by the following R code:

```
fit$svd
```

The output:

```
[1] 273.63958  99.57883
```


© insAnalytics

Potency Index– How the Variables Influencing the Functions

We calculate the Potency Index for all the variables, as below.

		LD1			LD2		
		Loadings	Square Loading	Potency Index	Loadings	Square Loading	Potency Index
Variables	Product Quality	2.762	7.629	10.405	4.855	23.574	88.353
Variables	Competetive Pricing	0.219	0.048	0.065	0.034	0.001	0.004
Variables	Speed of Delivery	1.998	3.991	5.444	9.039	81.696	306.196
Variables	Number of Complains	-6.002	36.027	49.138	2.292	5.252	19.684
	Eigen Value	273.640			99.579		
	Relative Eigen value	0.733			0.267		

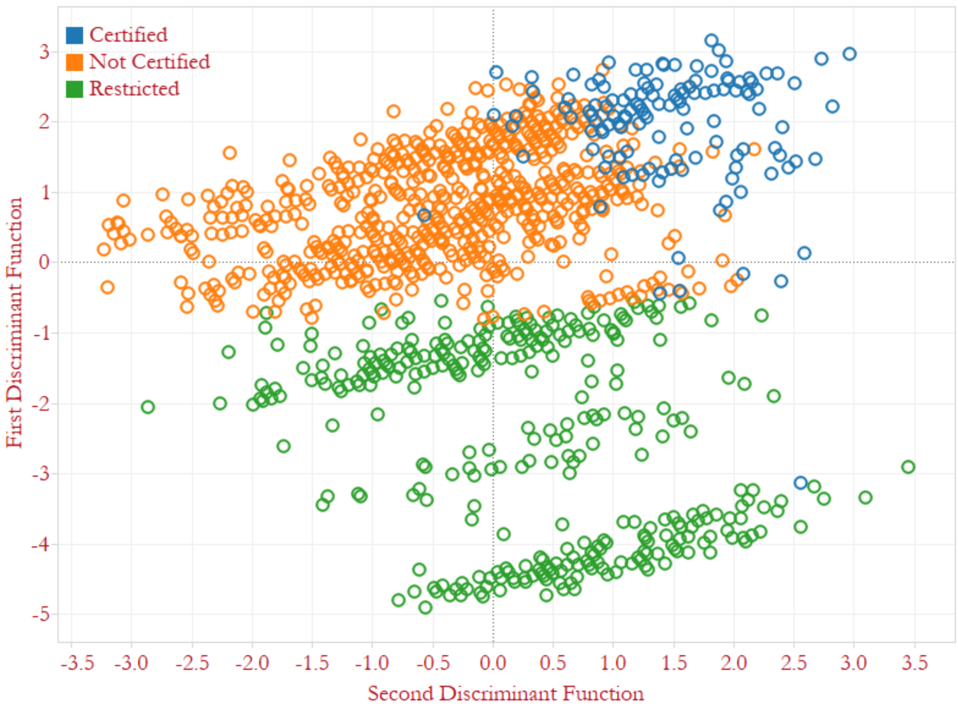
We find "Number of Complains" is more influencing LD1 where as LD2 is mostly influenced by "Speed of Delivery and "Product Quality"

© insAnalytics

Visualization – Territorial Plot

To look at the group with respect to the discriminant functions, we may try to create a territorial plot, using plot() function in R

A sample territorial plot using a random 2% of the data is shown below:



© insAnalytics

Classification Matrix– Measures of Accuracy

To measure the effectiveness, we calculate the classification matrix as below:

Right Decision

	Classified "Certified"	Classified "Not Certified"	Classified "Restricted"	Total
Actual "Certified"	4,863	1,954	89	6,906
Actual " Not Certified"	611	34,397	1	35,009
Actual "Restricted"		2,332	15,134	17,466
Total		38,683	15,224	59,381

Right Decision

Right Decision

© insAnalytics

Measuring Efficiency – Hit Ratio

We can calculate "**Accuracy Rate**" or "**Hit Ratio**" from the table.

	Classified "Certified"	Classified "Not Certified"	Classified "Restricted"	Total
Actual "Certified"	4,863	1,954	89	6,906
Actual " Not Certified"	611	34,397	1	35,009
Actual "Restricted"	0	2,332	15,134	17,466
Total	5,474	38,683	15,224	59,381

Hit Ratio = Correct Decision/ Total number of observations

= (4863+34397+15134)/59381

= **91.60%**

© insAnalytics

Discriminant Analysis

- Introduction ✓
- Assumptions and Terminologies ✓
- Discriminant Model Development Methodology ✓
- Building a Discriminant Analysis Model ✓
- Classification Using Discriminant Function ✓
- Classification with More than Two Categories ✓

© insAnalytics

Recapitulation & Key Takeaways

- **Discriminant analysis** is a classification technique where we classify observations into two or more categories using **mathematical function(s) of influencing variables**
- If the mathematical function is a linear function of predictors, it is called **linear discriminant analysis (LDA)**
- If the number of levels of response variables are more than two, then it is called **multivariate discriminant analysis (MDA)**
- The functions of predictors are called **discriminant functions** and the value of the function for a specific individual observation is called **discriminant score**
- Comparing the discriminant scores with **cut-off score values** we assign each observation into one of the response categories
- In discriminant analysis, all predictors has to be **numeric variable only**. They are assumed to follow **normal distribution** and **not correlated** among themselves
- The variables has be **homogeneous** (i.e. same variance) across levels of response variable
- One variable can be good predictor if
  - The mean of the variable is **significantly different** across response categories
  - The **between group Sum of Squares** of the variable is high
- **Potency Index** measure tells us the influence of one predictor in a classification
- **Trace value** indicates how many discriminant functions to be considered
- Once classified **confusion matrix** can be used to evaluate the performance of the classification rule.

© insAnalytics

Data Mining

Discriminant Analysis

© insAnalytics

© insAnalytics