United India Insurance Company

Business Understanding

Information about the Company

- United India Insurance Company is an Indian general insurance company, headquartered in Chennai, India. It was incorporated on 18 February 1938, and was nationalised in 1972.
- Previously it was a subsidiary of the General Insurance Corporation of India (GIC). But when GIC became a re-insurance company as per the IRDA Act 1999, its four primary insurance subsidiaries New India Assurance, United India Insurance, Oriental Insurance and National Insurance got autonomy.
- After nationalization, United India has 16385 nos. workforce spread across 2248 offices providing insurance cover to more than 10 million policyholders. The Company has a variety of insurance products to provide insurance cover from bullock carts to satellites
- It essentially runs as a Public Sector Undertaking(PSU).

Business Model

Cross-Selling Marketing Strategy:

Cross-selling is the action or practice of selling an additional product or service to an existing customer. In practice, businesses define cross-selling in many different ways. Elements that might influence the definition might include the size of the business, the industry sector it operates within and the financial motivations of those required to define the term.

Objective: of cross-selling can be either

1. To increase the income derived from the current clients
   or
2. To protect the relationship with the client or clients. The approach to the process of cross-selling can be varied.

- Unlike the acquiring of new business, cross-selling involves an element of risk that existing relationships with the client could be disrupted.
- For that reason, it is important to ensure that the additional product or service being sold to the client or clients enhances the value the client or clients get from the organization. In practice, large businesses usually combine cross-selling and up-selling techniques to increase revenue.

- For the vendor, the benefits are substantial. The most obvious example is an increase in revenue. There are also efficiency benefits in servicing one account rather than several.

- Most importantly, vendors that sell more services to a client are less likely to be displaced by a competitor. The more a client buys from a vendor, the higher the switching cost.

In the case of United India Insurance Company, there are two types of customers: Existing Customers and New Customers. Hence to maximise the benefits of cross-selling, existing customers are targeted.

For Existing Customers, three cases arise:

Case-1: **When New Coverage is greater than the Current Coverage**

- In this case, an existing customer who is already covered by one product insurance is offered an insurance coverage for another new product.
- The **current coverage** provided to a customer under a current product (1) may or (2) may not be equal to the coverage he/she receives under another new product
    1. Consider an example where a 30 year old customer's current product coverage is Rs. 10,00,000 and he is provided with Rs. 10,00,000 as coverage under another new product. The **New Coverage** for this customer is depicted as **Rs.20,00,000** i.e. the algebraic sum of the current product coverage and the coverage offered under another new product
    2. Consider an example where the same 30 year old customer's current product coverage is Rs. 10,00,000 and this time he is provided with Rs. 30,00,000 as coverage under another new product. The **New Coverage** for the same customer in this case is **Rs.40,00,000** i.e. the algebraic sum of the current product coverage and the coverage offered under another new product

Case-2: **When New Coverage is equal to the Current Coverage:**

- In this case, the validity or rather the lifetime period of a current product insurance under which a particular customer is covered is terminated or completed.
- Through the cross-selling strategy, the same customer, who is no longer covered under the current product insurance, is offered another new product insurance with a coverage equal to his/her expired or previous product coverage.
- Consider an example where the lifetime period of a particular customer's current product insurance ("ANS") with a current coverage of Rs. 10,00,000 has been completed or his current product insurance has expired.
- As part of cross-selling, the same customer is offered another new product insurance ("TLE") with a coverage equal to the previous expired product i.e. Rs 10,00,000. Hence, now the customer's **New Coverage** is Rs. 10,00,000

Case-3:**When New Coverage is less than the Current Coverage:**

- In this case, the validity or rather the lifetime period of a current product insurance under which a particular customer is covered is terminated or completed.

- Through the cross-selling strategy, the same customer, who is no longer covered under the current product insurance, is offered another new product insurance with a coverage not equal to his/her expired or previous product coverage.
- Consider an example where the lifetime period of a particular customer's current product insurance ("INV") with a **current coverage** of Rs. 40,00,000 has been completed or his current product insurance has expired.
- As part of cross-selling, the same customer is offered another new product insurance ("END") with a coverage equal to Rs 20,00,000. Hence, now the customer's **New Coverage** is Rs. 20,00,000.

Types of Cross-Selling: Broadly speaking, cross-selling takes three forms:

a) First, while servicing an account, the product or service provider may hear of an additional need, unrelated to the first, that the client has and offer to meet it.

b) Selling **add-on service**s - That happens when a supplier convinces a customer that it can enhance the value of its service by buying another from a different part of the supplier's company.

c) The third kind of cross-selling can be called selling a **solution**. In this case, the customer buying air conditioners is sold a package of both the air conditioners and installation services.

Benefits of Cross-Selling:

- Cross selling builds customer loyalty
- Strengthens the current customer relationship with the firm involved
- Increased sales revenue
- Improves customer and client satisfaction even in B2B businesses where business is conducted between companies rather than between a company and individual consumers
- Increases Customer Lifetime Value (CLV)

Customer Lifetime Value:

- The lifetime value of a customer, or customer lifetime value (CLV), represents the total amount of money a customer is expected to spend in a business, or on its products, during his/her lifetime.
- To calculate customer lifetime value you need to calculate average purchase value of a customer, and then multiply that number by the average purchase frequency rate to determine customer lifetime value.

Services:

- Motor Insurance
- Health Insurance
- Travel Insurance

- Personal Accident Insurance
- Householder’s Insurance
- Shopkeeper’s Insurance
- Fire,Marine,Industry,Liability,Micro and Credit Insurance

Data Understanding/Description

The data provided by the insurance company is spread across one dataset with 100000 records and 16 variables; the contents of which are:

- **Age:** Listing Customers’ age ranging from 18 years to 60 years
- **Gender**: Listing whether the customer is a male or a female
- **Marital Status**: Listing whether a customer is married, divorced or single
- **Family Members:** Listing the number of members in the family of a customer i.e. ranging from 1 to 10 members
- **Education:** Listing the education background of customers
- **Occupation and Job Title:** Listing the occupations and job titles of customers
- **Income**: Listing the income of customers income in Lakhs of Rupees per Annum
- **Current Product:** Listing whether a customer has currently chosen a product insurance or not i.e. it indirectly implies whether a customer is new or already existing.
- **Current Product Type:** Listing the product type (code) of a current product
- **Current Coverage**: Listing the amount (in rupees) current product insurance covers under the current policy or scheme
- **New Product Type:** Listing the product type (code) of a new product
- **New Coverage**: - If a customer is already provided coverage for a current product and opts for a new product: **New Coverage** exhibits the cumulative total coverage(in rupees) for both the products chosen by a customer
- **Rating**: Listing the rating given by customers as Cold, Warm or Hot for the services provided by the insurance company
- **Converted**: Listing whether the cross-selling strategy of the insurance company has been successful or not i.e. whether the new and existing customers have converted or not
- **Status**: Listing whether the cross-selling strategy of the insurance company has been successful or not i.e. whether the new and existing customers have converted or not.
  If customers have not converted then various **codes** have been allocated by the company based on the insurance policy of a customer

Data Audit

Data Audit involves the following steps:

1. Formatting changes – Examining variables that need to be converted to an appropriate data type before analyzing them

2. Validation – Finding out whether the variables contain any missing values or not
3. Checking for Errors – Searching for Variables that contain any outliers or negative values
4. Subsetting – Removal of unnecessary or insignificant variables from the data set

Age:

- Missing Data Present - 44 missing records
- Ranging from 18 to 60
- No presence of any outliers
- Requires Data cleaning
- The Age Variable was converted from a double data type to a numeric data type
- Missing Values were imputed by considering the minimum value between the mean and median values which were calculated excluding the missing data
- In this case Median (39 years) was found to be lesser than Mean. Hence Median was used to impute missing values

Gender:

- Missing Data Present - 182 Missing Records
- Requires Data Cleaning
- The Gender Variable was converted to a character data type
- Females-39893 , Males-59925
- Missing Values were imputed with the mode(Male) calculated for the variable
- After Imputation of missing values, datatype was converted to factor.

Marital Status:

- Missing Data Present - 42 Missing Records
- Requires Data Cleaning
- Data type: Character
- Divorced-23160 , Married-53950 , Single-22848
- Missing Values were imputed with the mode(Married) calculated for the variable
- After Imputation of missing values, datatype was converted to factor.

Family Members:

- Missing Data Present - 22 Missing Records
- Ranging from 1 to 10
- Requires Data Cleaning
- No presence of any outliers
- The Age Variable was converted from a double data type to a numeric data type

- Missing Values were imputed by considering the minimum value between the mean and median values which were calculated excluding the missing data
- Mean-4.659 , Median-4.0 (excluding Missing Values). Hence Median was used to impute missing values

Education:

- Missing Data Present - 46 Missing Records
- Requires Data Cleaning
- Data type: Character
- BD-29586 LHS-13399 MD- 22543 NE-7232 PD-14746 UHS-12448
- Missing Values were imputed with the mode calculated for the variable
- After Imputation of missing values, datatype was converted to factor.

Occupation:

- Missing Data Present - 46 Missing Records
- Requires Data Cleaning
- Data type: Character
- SE-39443 SFT-8522 SPT- 51992
- Missing Values were imputed with the mode calculated for the variable
- After Imputation of missing values, datatype was converted to factor.

Job Title:

- Missing Data Present - 123 Missing Records
- Requires Data Cleaning
- Data type: Character
- BA-7980 CB-5396 CF-2142 DD-9795 FH-23780 OC-1660 OM-5935
- OT-6530 PA-1609 PG-19623 PR-7666 RR-7761
- Missing Values were imputed with the mode calculated for the variable
- After Imputation of missing values, datatype was converted to factor.

Income:

- Missing Data Present - 22 Missing Records
- Ranging from 1 to 10
- Requires Data Cleaning
- No presence of any outliers
- The Income Variable was converted from a double data type to a numeric data type
- Missing Values were imputed by considering the minimum value between the mean and median values which were calculated excluding the missing data
- Mean-4.96 , Median-3.0 (excluding Missing Values). Hence Median was used to impute missing values

Current Product:

- Missing Data Present - 18 Missing Records
- Requires Data Cleaning
- Data type: Character
- No-42862   Yes-57120
- Missing Values were imputed with the mode calculated for the variable
- After Imputation of missing values, datatype was converted to factor.

Current Product Type:

- Missing Data Present - 33 Missing Records
- Requires Data Cleaning
- Data type: Character
- ANS-18461 END-5487 INV-10346 NA-42855 PMT-8477 TLE-14341
- Missing Values were imputed with the mode calculated for the variable
- After Imputation of missing values, datatype was converted to factor.

Current Coverage:

- Missing Data Present - 49 Missing Records
- Ranging from 0 to 15000000
- Requires Data Cleaning
- Positive Outliers are Present
- The Current Coverage Variable was converted from a character data type to a numeric data type
- Missing Values were imputed by considering the minimum value between the mean and median values which were calculated excluding the missing data
- Mean-3633770 , Median-50000 (excluding Missing Values). Hence Median was used to impute missing values

New Product Type:

- Missing Data Present - 47 Missing Records
- Requires Data Cleaning
- Data type: Character
- ANS-29151 END-12718 INV-20429 PMT-20429 TLE-25067
- Missing Values were imputed with the mode calculated for the variable
- After Imputation of missing values, datatype was converted to factor.

New Coverage:

- Missing Data Present - 127 Missing Records

- Ranging from 1000000 to 15000000
- Requires Data Cleaning
- No Outliers Present
- The New Coverage Variable was converted from a character data type to a numeric data type
- Missing Values were imputed by considering the minimum value between the mean and median values which were calculated excluding the missing data
- Mean-6106250 , Median-3000000 (excluding Missing Values). Hence Median was used to impute missing values

Rating:

- Missing Data Present - 44 Missing Records
- Requires Data Cleaning
- Data type: Character
- Cold-47551 Hot-22988 Warm-29417
- Missing Values were imputed with the mode calculated for the variable
- After Imputation of missing values, datatype was converted to factor.

Converted:

- Missing Data Present - 49 Missing Records
- Data type: Character
- Converted- 38296   Not Converted -61655

Status:

- No Missing Data Present
- To convert this variable to a target variable convenient for data modelling, all the data points written as codes are replaced with “Not Converted”
- The new variable called “**Status_new**” can now be used as a target variable for application of predictive modelling as it is a categorical variable with no missing values.
- **Status_new** was converted from a character data type to a factor data type.

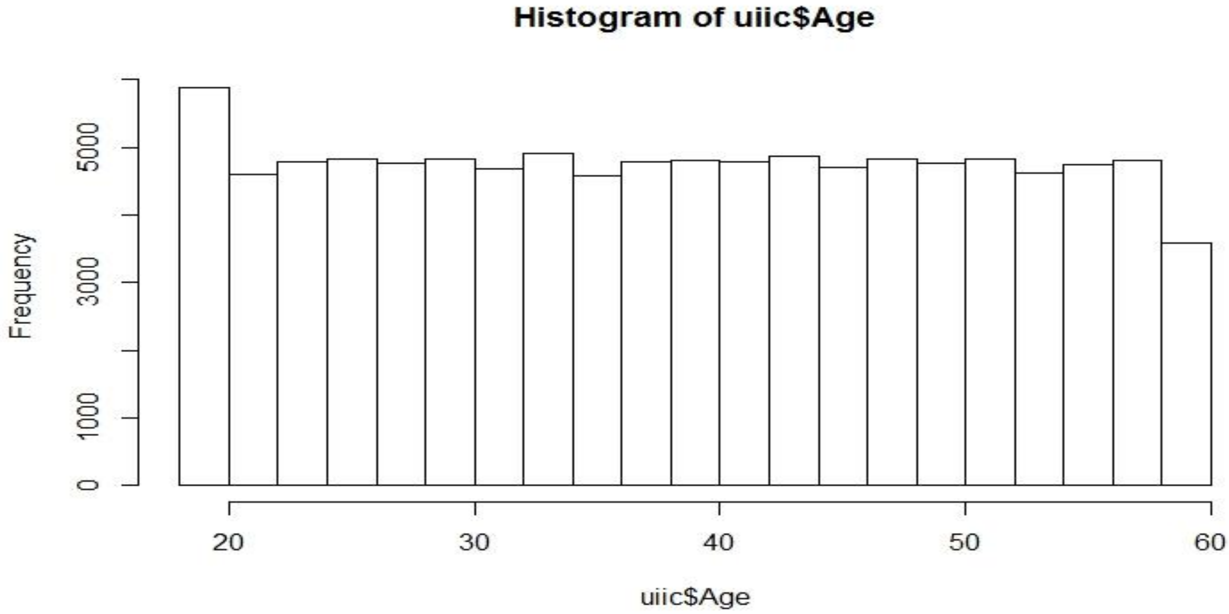
Data Exploration

Age:

- Minimum Age - 18 years , Maximum Age - 60 years
- Average Age of Customers is 39 years
- Maximum no. of Customers (2480) fall in the category of 33 years and 51 years

- Customers who are either 60 years of age (1189) or 18 years of age (1177) are the least when compared to customers of other ages
- All other age categories have about 2300-2400 customers each

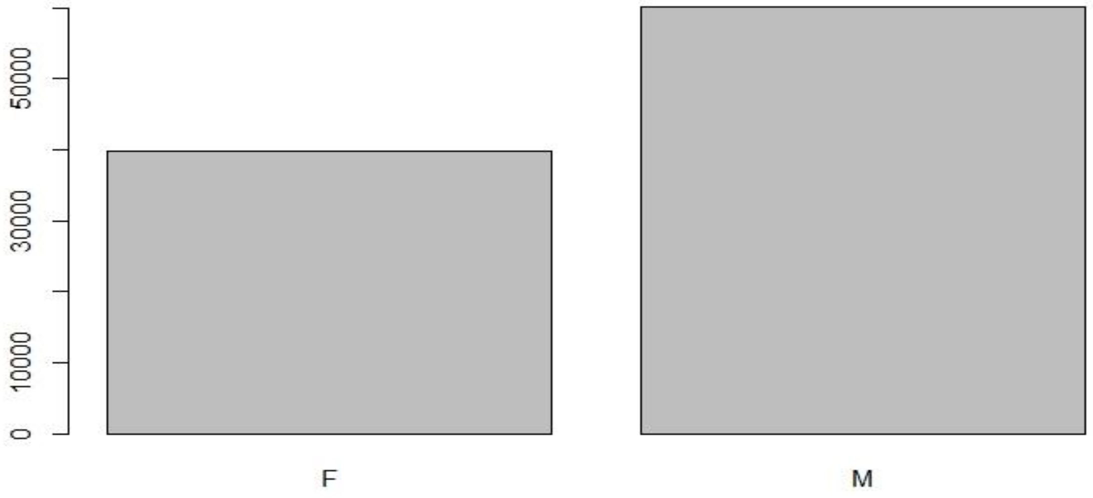
Histogram showing Number of Customers (Y-Axis) vs Age of Customers (X-Axis)



Gender:

- Out of 1,00,000 customers- 39,893 are Females and 60,107 are Males
- Among Females, those who are aged 51 years (1004) are maximum in number when compared to other age groups.
- Among Males,those who are aged 50 years (1512) are maximum in number when compared to other age groups

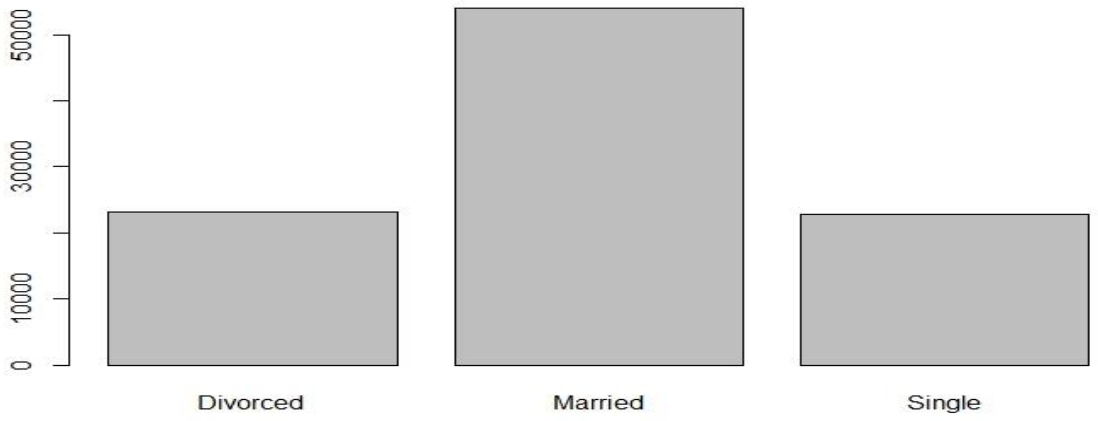
Bar Plot showing Number of Customers (Y-Axis) vs Gender of Customers (X-Axis)



Marital Status:

- Out of 1,00,000 customers: Divorced-23160 Married-53992 Single-22848
- 21,519 Females and 32473 Males are Married
- 9075 Females and 13773 Males are Single
- A Total of 5,602 Customers who belong to category of 19-21 years are Single
- None of the Customers who belong to the age group 18-21 years are Divorced

Bar Plot showing Number of Customers (Y-Axis) vs Marital Status of Customers (X-Axis)



Family Members:

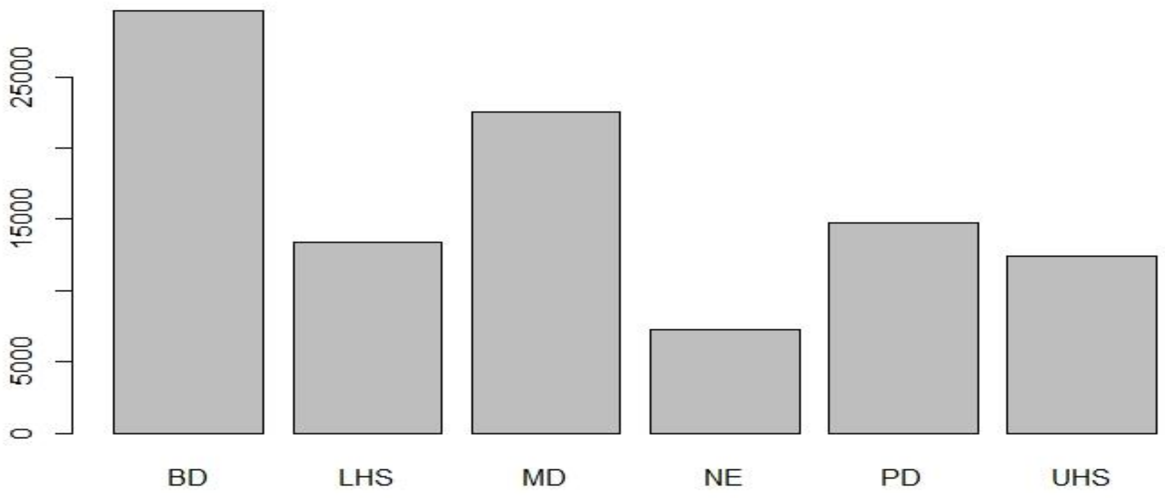
- Out of 1,00,000 customers: 1 Member-8654 2 Members-19011
- 3 Members-11795 4 Members-13322 5 Members-11734 6 Members-10376
- 7 Members-7414 8 Members- 5850 9 Members-7487 10 Members- 4357
- 118 Customers who are aged 34 have 10 members each in their families
- 373 Customers who are aged 21 have only one member each in their families
- Most Customers have at least 2 members each in their families
- 11929 Customers who are Married have only 2 members each in their families
- Only 2 Customers who are Married have only one other member in their families

Education:

- Out of 1 Lakh Customers:
- Bachelor's Degree-29586 LHS-13399 Master’s Degree-22543 NE-7232 Professional Degree-14746 UHS-12448
- 1207 Customers aged 19 years are not enrolled in any academic program
- Only 2 twenty year olds and 1 eighteen year old have a bachelor’s degree
- None of the Customers belonging to 18-20 years age group have either a master’s degree or a professional degree
- Among Males, 17863 have a Bachelor’s Degree and 13592 have a Master’s Degree and 4309 Males are not enrolled or don't have a degree.
- Among Females, 11769 have a Bachelor’s Degree and 8951 have a Master’s Degree

- Among Married Customers, 16868 have a Bachelor’s Degree and 12653 have a Master’s Degree
- 905 Divorced customers are not enrolled in any academic program or do not have any educational qualification

Bar Plot showing Number of Customers (Y-Axis) vs Education Background of Customers (X-Axis)



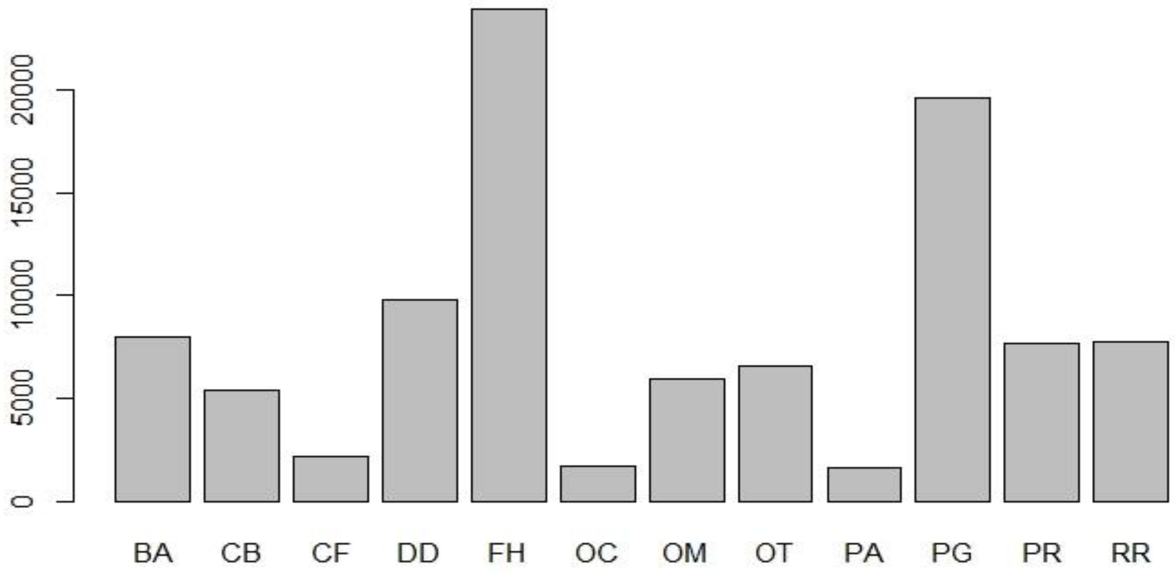
Occupation:

- Out of 1 Lakh Customers: SE-39443 , SFT-8522 and SPT-52035
- Among Females, SE-15731, SFT-3458 and SPT-20704
- Among Males, SE-23712 , SFT-5064 and SPT-31331
- Among Divorced Customers: SE-9483 SFT-1665 SPT-12012
- Among Married Customers: SE-21887 SFT-4228 SPT-27877
- Among Single Customers: SE-8073 SFT-2629 SPT-12146
- Among Customers who have a Bachelor’s Degree: SE-12310 SFT-5 SPT-17317
- Among Customers who have their Education as LHS, only 4 have their Occupation as SPT
- Among Customers who have their Education as an MD,NE,PD or UHS, none have their Occupation as SFT
- Only 4 Customers who have their qualification as LHS, have their occupation as SPT

Job Title:

- Out of 1,00,000 Customers:
- BA-7980 CB-5396 CF-2142 DD-9795 FH-23903 OC-1660 OM-5935
- OT-6530 PA-1609 PG-19623 PR-7666 RR-7761
- Among the Female Customers, most work as a FH (9517) and 7762 work as a PG
- Among the Male Customers, most work as a FH (14386) and 11861 work as a PG
- Among the Customers who have a Bachelor's Degree: 10,629 work as a FH, 3 work as an OT, 2 work as a CB, 1 works as an OC while none of them work as a CF,OM or as a PA
- Among Customers who have their Qualification as LHS: 3948 of them work as a DD while none of them work as a BA,PG,PR or as a RR
- Among the Customers who have a Master's Degree, 8,052 of them work as a FH while none of them work as CB,CF,DD,OC,OM,OT or as a PA
- Among the Customers who have a Professional Degree, 5196 work as a FH while none of them work as a BA,PG,PR or as a RR
- Among the Customers who have their occupation as SE; 23831 work as a FH, while none of them work as a BA,OC,PA,PR or RR
- Among the Customers who have their occupation as SFT; 2626 work as a DD, while none of them work as a BA,CF,PG,PR or RR
- Among the Customers who have their occupation as SPT; Most work as a PG, while none of them work as a CF

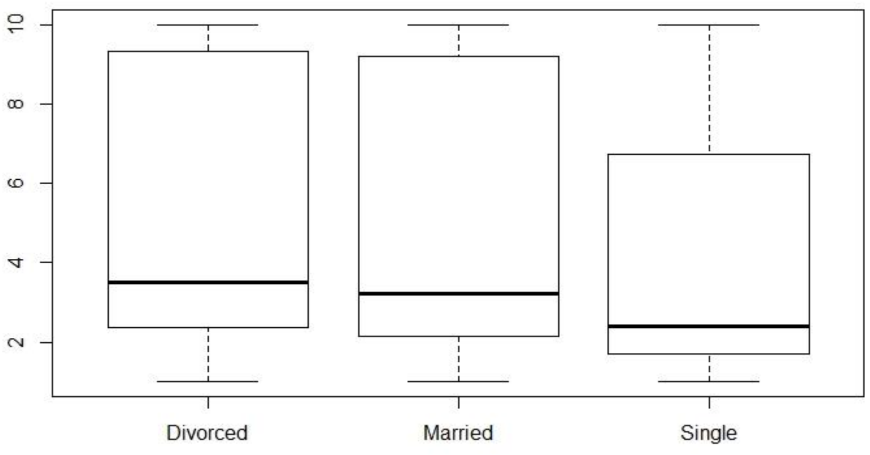
Bar Plot showing Number of Customers (Y-Axis) vs Job Title of Customers (X-Axis)



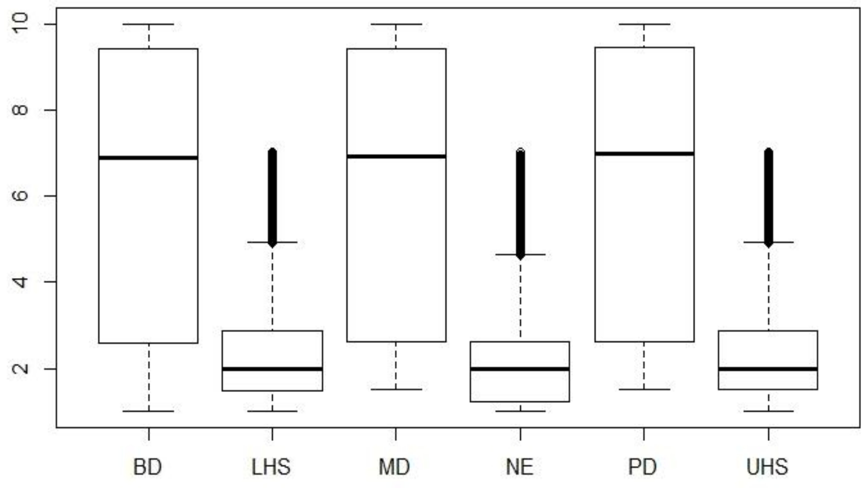
Income:

- Minimum Income of any Customer is One Lakh Rupees per annum; while Maximum is Ten Lakh Rupees per annum.
- Average Income is nearly 5 Lakhs Per Annum among all Customers
- Average income for Customers aged 18 years is 1.83 Lakhs per Annum and people aged 34 years and above earned at least an average of 5 Lakhs Per Annum
- Males have a higher income than Females in general but the average income for both groups is nearly equal (4.9 Lakhs Per Annum)
- Although Married Customers earn more than Divorced Customers in general, their average salary is slightly less than the divorced customers.
- The average income of Customers with a Bachelor’s, Master’s or Professional degree is nearly equal i.e. approximately 6.2 Lakhs per Annum; Customers with other educational qualifications earn nearly 2.35 Lakhs Per Annum
- Average Income of customers whose occupation is SPT is 6.89 Lakhs per annum; this is higher than than the average income of customers with other occupations

Box Plot showing Income of Customers (Y-Axis) vs Marital Status of Customers (X-Axis)



Bar Plot showing Income of Customers (Y-Axis) vs Educational Qualification of Customers (X-Axis)

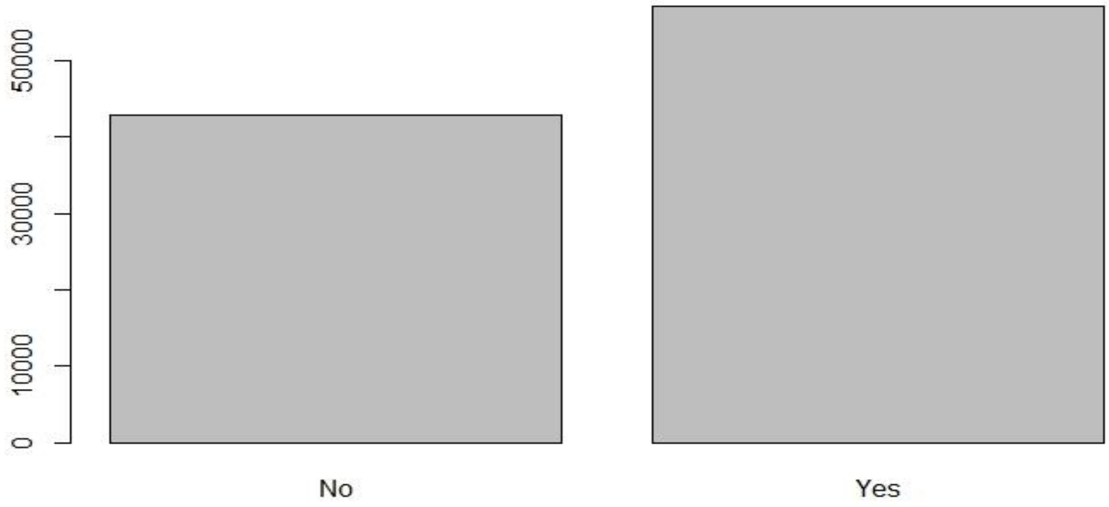


Current Product:

- Out of 1 Lakh customers, 57,129 are existing customers i.e. who are already covered through a product insurance and 42,871 are new customers i.e. who are not covered through any kind of product insurance.
- Out of all the existing customers, 17174 are females and 25697 are males

- Out of all the new customers, 22719 are females and 34410 are males
- Out of all the existing customers: 32437 are married, 9253 are single and the rest divorced
- Out of all the new customers: 21555 are married, 13595 are single and the rest divorced
- Among both existing and new customers, most customers have a bachelor’s degree as their academic qualification
- For Existing Customers, both cumulative income (Algebraic sum of incomes of all existing customers) and average income are greater than those of their new counterparts
- Average income of existing customers is 5.2 Lakhs per Annum and for new customers it is 4.64 Lakhs per Annum

Bar Plot showing Number of Customers (Y-Axis) vs Whether a Customer is an Existing Customer (X-Axis)



Current Product Type:

- Among all the existing customers: 18,461 have chosen “ANS” , 5,487 have chosen “END” , 10,346 have chosen “INV”, 8,477 have chosen “PMT” and 14,341 have chosen “TLE”

- Cumulative income (algebraic sum of incomes of customers with ANS product) of customers who have "ANS" product insurance is higher than all other cumulative incomes of customers with other product insurances

Current Product Type and Average Income (in Lakhs of rupees per annum) of Existing Customers:

ANS	END	INV	PMT	TLE
5.78	4.70	6.74	3.89	4.31

Current Coverage:

- The minimum current coverage for any customer is Rs.0 while the maximum current coverage offered to a customer is Rs. 1,50,00,000

Current Product Type and Average Current Coverage offered (in Lakhs of rupees) for Existing Customers:

ANS	END	INV	PMT	TLE
66	63	69	58.7	59.6

New Product Type:

- Among all the existing and new customers: 29198 have chosen "ANS" , 12718 have chosen "END" , 20429 have chosen "INV", 12588 have chosen "PMT" and 25067 have chosen "TLE"

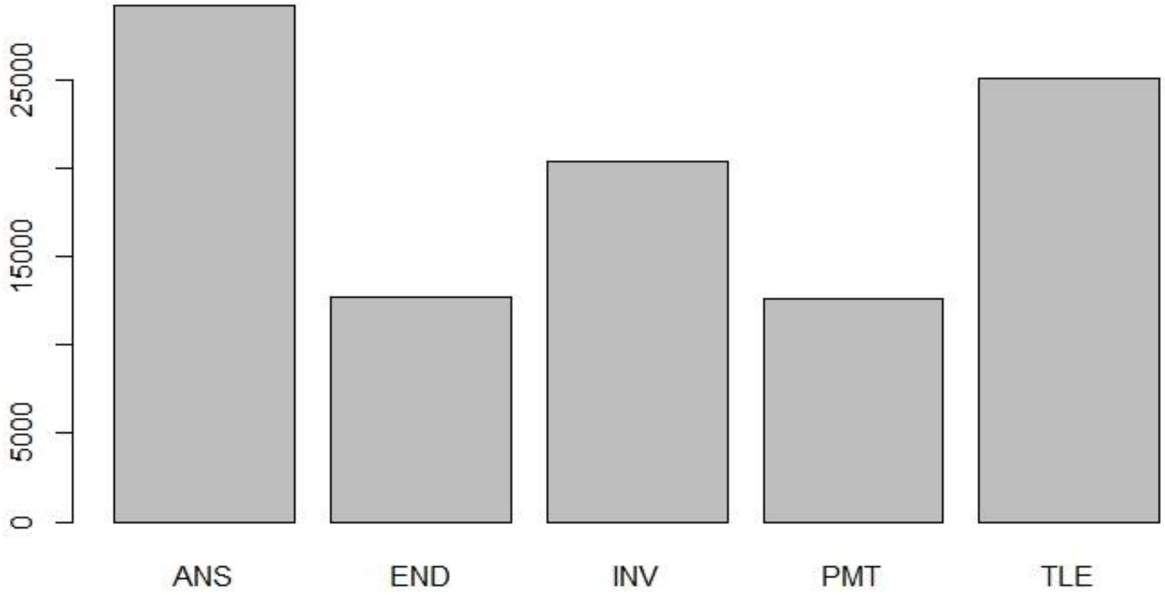
Number of Existing and New Customers for Each New Product Type:

	ANS	**END**	**INV**	**PMT**	**TLE**
Number of New Customers	10827	7785	9930	6135	8194

Number of Existing Customers	18371	4933	10499	6453	16873

	Customers With ANS as New Product Type	**Customers With END as New Product Type**	**Customers With INV as New Product Type**	**Customers With PMT as New Product Type**	**Customers With TLE as New Product Type**
Customers with ANS as Current Product Type	7	862	5090	2211	10291
Customers with END as Current Product Type	1103	0	2766	783	835
Customers with INV as Current Product Type	4411	2726	0	1356	1853
Customers with PMT as Current Product Type	2688	728	1169	0	3892
Customers with TLE as Current Product Type	10154	616	1472	2099	0

Bar Plot showing Number of Customers (Y-Axis) vs New Product Type of Customers (X-Axis)



New Coverage:

- The minimum new coverage for any customer is Rs.10,00,000 while the maximum new coverage offered to a customer is Rs. 1,50,00,000

- **New Product Type and the Average New Coverage offered (in Lakhs of rupees) for all Customers**:

ANS	END	INV	PMT	TLE
60.45	61.57	61.24	64.34	59.72

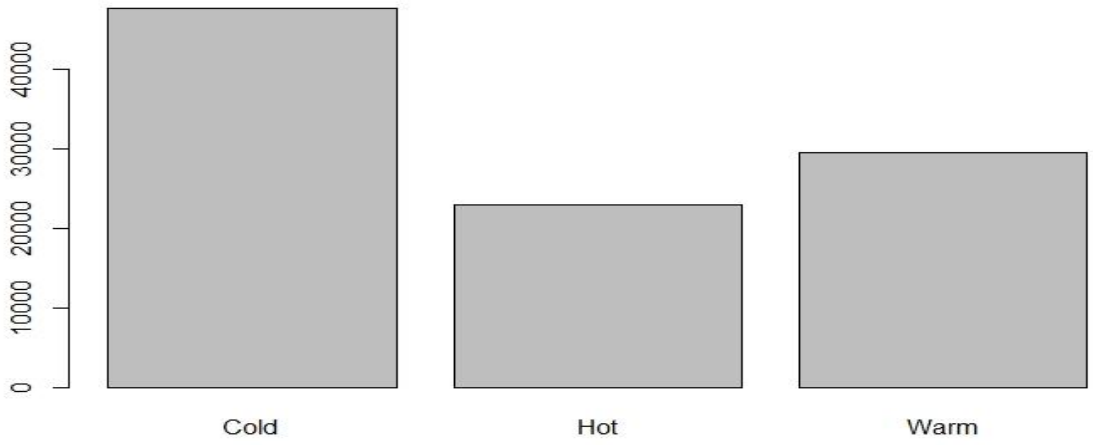
Rating:

- Out of all the customers: 47,595 customers have rated their experience as “Cold”, 29,417 have rated their experience as “Warm” and 22,988 customers have rated their experience as “Hot”

	Cold Rating	Warm Rating	Hot Rating
Number of Females	19103	11646	9144
Number of Males	28492	17771	13844

	Cold Rating	Warm Rating	Hot Rating
Number of New Customers	19808	11314	11749
Number of Existing Customers	27787	18103	11239

Bar Plot showing Number of Customers (Y-Axis) vs Rating Given by Customers (X-Axis)



Status_New:

- Out of all the 1 lakh Customers, 38317 converted i.e. the cross-selling strategy was successful.
- While 61,683 customers did not convert i.e. the cross-selling strategy was not successful.

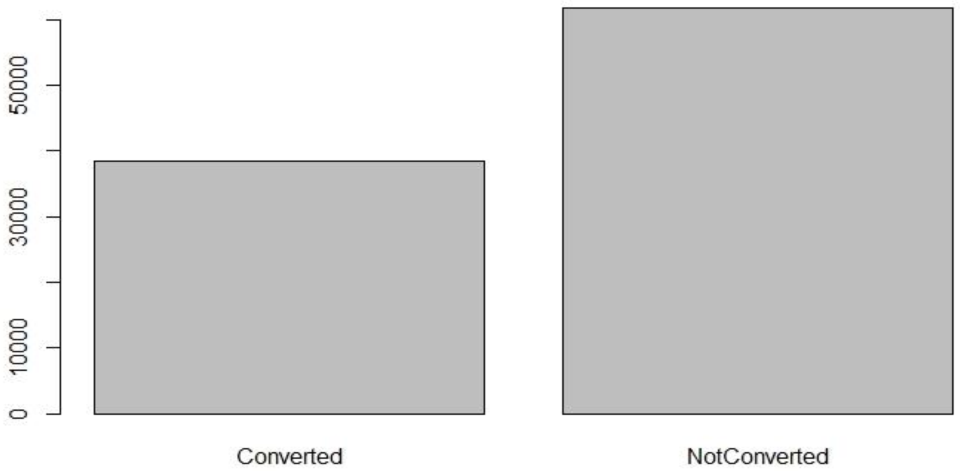
	Customers who Converted	**Customers who Did Not Convert**
Number of Female Customers	15287	24606
Number of Male Customers	23030	37077

	Customers who Converted	**Customers who Did Not Convert**
Average Income of Customers (in Lakhs of Rupees per Annum)	7.50	3.40

	Customers who Converted	**Customers who Did Not Convert**
Number of New Customers	18955	23916
Number of Existing Customers	23916	37767

	Customers who Converted	**Customers who Did Not Convert**
Customers who rated as Cold	2318	45277
Customers who rated as Warm	15341	14076
Customers who rated as Hot	20658	2330

Bar Plot showing Number of Customers (Y-Axis) vs Status of Customers (X-Axis)



Model Building:

- To Predict whether a customer converted or did not convert i.e. whether the cross-selling strategy of the company was successful or not , three predictive modelling techniques were applied.

Namely:

- Logistic Regression
- Decision Tree
- KNN(‘K’- Nearest Neighbours) Algorithm

Logistic Regression :

- Logistic regression is used to explain the relationship between a binary dependent variable and one or more discrete, categorical or continuous independent variables.

Steps in Performing Logistic Regression:

1. Identify the required data that is suitable for the model.
2. Check for any high positive/Negative skewness in given independent variables and apply transformations if necessary.
3. Check whether all the independent variables are following a linear relationship with the dependent variable.
4. Build a simple linear regression model(s) and check significance and coefficient of determination of the linear relation.
5. Perform Data partition and divide the data into train and test using Machine Learning

- The Data is divided in such a way that the Train data consists of 75% (75,000 Records) of the Data and the Test data contains only 25% (25,000 Records) of the Data

6. Use train data and build a multiple logistic regression model and validate the assumptions of the model.
7. Identify and eliminate the multicollinear variables from the model.
   - Multicollinearity exists when two or more independent variables represent an approximate or exact linear relationship with respect to one another.
   - All variables with an **Variance Inflation factor** (VIF) value greater than 3.5 are considered to be Multicollinear Variables and are removed before modelling
   - The variable "Income" was found to be Multicollinear i.e. its VIF value was found to be greater than 3.5.
8. Once the data is free of Multicollinearity, insignificant independent variables are suppressed manually or by using best regression line techniques namely: **Forward Stepwise Regression** and **Backward stepwise Deletion.**
   - A variable is considered to be insignificant when its "P" value exceeds the significance level i.e. set as 0.05 in this case

     1) **Forward Stepwise Regression:**
        - This technique starts with the single best variable and adds more variables to build the model into a more complex form.
        - After Forward Stepwise Regression, no insignificant variables were removed for modelling.
        - Null Deviance was 99920, Residual Deviance was 50700 and AIC was 50730.

     2) **Backward Stepwise Deletion:**
        - This technique considers all variables initially and then the model is reduced by removal of insignificant variables until only significant variables are left.
        - After Backward Stepwise Deletion, few insignificant variables were removed.
        - The insignificant variables that were removed are: Gender, \`Marital Status\`, \`Family Members\`,\`Current Product\` and \`Current Product Type\`.
        - Null Deviance was 99920, Residual Deviance was 50700 and AIC was 50720.

9. Finalize the appropriate model by selecting the best regression line technique where the model has the least AIC value.

10. Use Test data to check the prediction of the model.
11. Validate the model in prediction using methods like confusion matrix and area under the ROC curve.

Logistic Regression Line:

```
Call:  glm(formula = Status_new ~ Age + Education + Occupation + `Job Title` +
   `Current Coverage(In Rs)` + `New Product Type` + `New Coverage(In Rs)` +
   Rating, family = "binomial", data = trainer)

Coefficients:
          (Intercept)                Age              Education
           4.044e+00             -3.334e-02            3.515e-02
          Occupation              `Job Title`    `Current Coverage(In Rs)`
          -2.734e-01             -3.218e-02            1.288e-07
    `New Product Type`  `New Coverage(In Rs)`      Rating
           2.157e-02              2.345e-08           -2.729e+00
```

- Among both the techniques, the regression line technique in which the **Akaike Information Factor (AIC), Null Deviance and Residual Deviance** values are minimum, is considered for predictive modelling.
- **AIC** is is an estimator of the relative quality of statistical models for a given set of data. It estimates the quality of each model, relative to each of the other models.
- **Null Deviance** shows how well the dependent variable is predicted by a model that includes only the intercept.

- For Validating the Logistic Regression Model, two techniques were used:

a) **Confusion Matrix:**

- In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one.
- Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class.

	Predicted Class -**‘Converted’**	Predicted Class-**’Not Converted’**
Actual Class- **‘Converted’**	8191 True Positives	1300 False Positives
Actual Class- **‘Not Converted’**	1571 False Negatives	13938 True Negatives

b) **ROC Curve:**

- In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (Specificity) for different cut-off points.
- Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.
- Area under the Curve was found to be **0.9184882** i.e. higher the area under curve, greater the predictive accuracy of the model.

Decision Tree :

- Decision tree is a graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions.
- It is mostly used in Machine Learning and Data Mining applications using R.

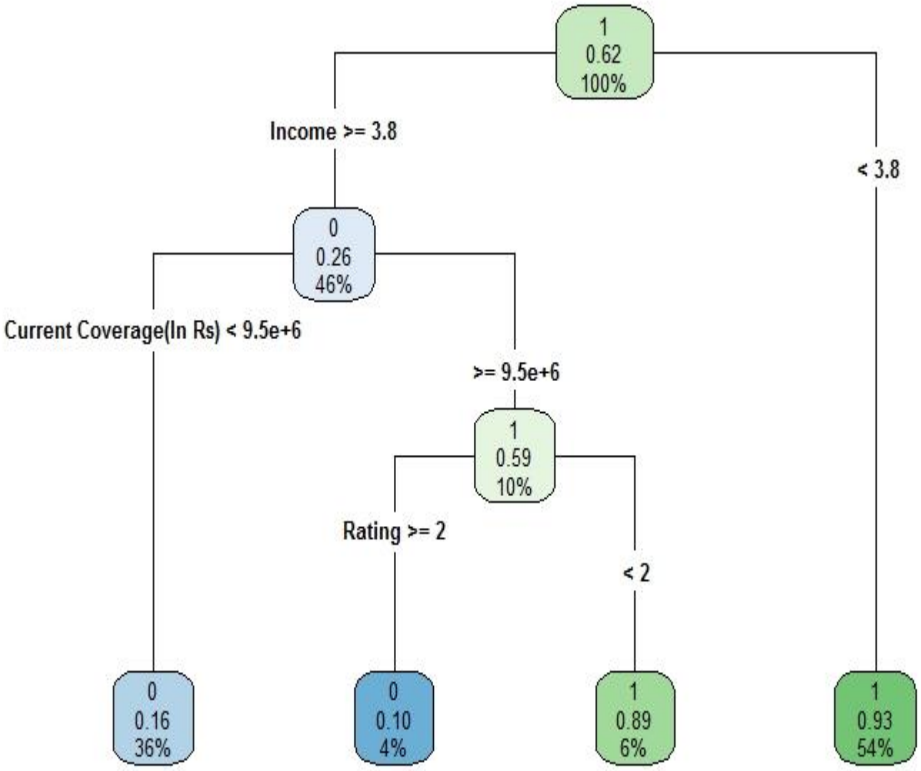
Steps in Creating a Decision Tree Model:

1. Identify the required data that is suitable for the model.
2. Check for any high positive/Negative skewness in given independent variables and apply transformations if necessary.
3. Check whether all the independent variables are following a linear relationship with the dependent variable.
4. Perform Data partition and divide the data into train and test using Machine Learning

- The Data is divided in such a way that the Train data consists of 75% (75,000 Records) of the Data and the Test data contains only 25% (25,000 Records) of the Data

5. Use train data and build a simple decision tree model and validate the assumptions of the model.
6. Identify and eliminate the multicollinear variables from the model.
    - Multicollinearity exists when two or more independent variables represent an approximate or exact linear relationship with respect to one another.
    - All variables with an Variance Inflation factor (VIF) value greater than 3.5 are considered to be Multicollinear Variables and are removed before modelling
7. Once the data is free of Multicollinearity, insignificant independent variables are suppressed using Machine Learning.
8. The most significant variables absolutely necessary for building a Decision Tree were found to be Income, Current Coverage and Rating.
9. Pruning is then performed on the decision tree using a **Complexity Parameter** to avoid overfitting in the tree.
10. The **Complexity Parameter** (CP) is used to control the size of the decision tree and to select the optimal tree size. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then tree building does not continue.

    - The CP value corresponding to the least relative and cross validation errors was chosen and found to be 0.01. Pruning was not effective as both the original and pruned models were found to be identical. The obtained decision tree:



Business Rules for Decision Tree:

- 10% of all the customers i.e. 10,000 customers **converted** through the cross-selling scheme when their Income was greater than or equal to Rs. 3.8 Lakh per Annum, their Current Coverage was greater than or equal to Rs. 95 Lakh and when the Rating was "HOT"
- 16% of all customers i.e. 16,000 customers **converted** through the cross-selling scheme when their Income was greater than or equal to Rs. 3.8 Lakh per Annum and their Current Coverage was less than Rs. 95 Lakh

11. Use Test data to check the prediction of the model.

12. Validate the model in prediction using methods like confusion matrix

- For Validating the Decision Tree Model, one technique was used:

a) Confusion Matrix:

	Predicted Class -**'Converted'**	Predicted Class-**'Not Converted'**
Actual Class- **'Converted'**	8280 True Positives	1211 False Positives
Actual Class- **'Not Converted'**	1569 False Negatives	13940 True Negatives

K- Nearest Neighbours Algorithm :

- In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression.[1] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:
- In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its "k" nearest neighbors.

Steps in Creating a K-Nearest Neighbours Model:

1. Identify the required data that is suitable for the model.
2. Check for any high positive/Negative skewness in given independent variables and apply transformations if necessary.
3. Check whether all the independent variables are following a linear relationship with the dependent variable.
4. The categorical target variable is excluded from the dataset and then the data is standardized through scaling.
5. Once the standardization is performed, the target variable is added back to dataset.
6. Perform Data partition and divide the data into train and test using Machine Learning

- The Data is divided in such a way that the Train data consists of 75% (75,000 Records) of the Data and the Test data contains only 25% (25,000 Records) of the Data

  7. Use train data and build a simple k-NN model and validate the assumptions of the model.

8. Identify and eliminate the multicollinear variables from the model.

- Multicollinearity exists when two or more independent variables represent an approximate or exact linear relationship with respect to one another.
- All variables with an Variance Inflation factor (VIF) value greater than 3.5 are considered to be Multicollinear Variables and are removed before modelling

9. Once the data is free of Multicollinearity, insignificant independent variables are suppressed using Machine Learning.

- The “k” value, for which the “Kappa” index/value is maximum, is considered the optimum “k” (number of nearest neighbours) value.
- **“Kappa”** index gives the accuracy of classifying records.
- In this case, the number of nearest neighbours is 9 (k=9).

10. Use Test data to check the prediction of the model.

11. Validate the model in prediction using methods like confusion matrix

- For Validating the Model created through the K-NN Algorithm, one technique was used:

a) Confusion Matrix:

	Predicted Class **-‘Converted’**	Predicted Class-**’Not Converted’**
Actual Class**- ‘Converted’**	9951 True Positives	1710 False Positives
Actual Class**- ‘Not Converted’**	1544 False Negatives	16794 True Negatives

Results:

- Based on the accuracies obtained for all three modelling techniques, the model developed through the **K-Nearest Neighbours Algorithm** has the highest accuracy i.e. **89.15297 %** calculated through validation by a confusion matrix.
- The Area under the Curve for the model created through the K-NN Algorithm is **0.8866341** i.e. higher the area under curve, greater the predictive accuracy of the model.

Cross Validation:

<table>
<tr><th></th><th>Logistic Regression Model</th><th>Decision Tree Model</th><th>K- Nearest Neighbours Model</th></tr>
<tr><td><b>Confusion Matrix</b></td>
<td><table>
<tr><th></th><th>Predicted Class - <b>‘Converted’(0)</b></th><th>Predicted Class- <b>’Not Converted’ (1)</b></th></tr>
<tr><td>Actual Class- <b>‘Converted’(0)</b></td><td>8191</td><td>1300</td></tr>
<tr><td>Actual Class- <b>‘Not Converted’(1)</b></td><td>1571</td><td>13938</td></tr>
</table></td>
<td><table>
<tr><th></th><th>Predicted Class - <b>‘Converted’(0)</b></th><th>Predicted Class- <b>’Not Converted’ (1)</b></th></tr>
<tr><td>Actual Class- <b>‘Converted’ (0)</b></td><td>8280</td><td>1211</td></tr>
<tr><td>Actual Class- <b>‘Not Converted’ (1)</b></td><td>1569</td><td>13940</td></tr>
</table></td>
<td><table>
<tr><th></th><th>Predicted Class - <b>‘Converted’(0)</b></th><th>Predicted Class- <b>’Not Converted’ (1)</b></th></tr>
<tr><td>Actual Class- <b>‘Converted’ (0)</b></td><td>9951</td><td>1710</td></tr>
<tr><td>Actual Class- <b>‘Not Converted’ (1)</b></td><td>1544</td><td>16794</td></tr>
</table></td></tr>
<tr><td><b>Predictive Accuracy</b></td><td>88.516%</td><td>88.88%</td><td>89.153%</td></tr>
</table>