BADM Project Report

PGPBM 2019-21

****

**AST Insurance Company**

*by*

**Group 3**

Gnana C P K – 1903012

Gurunathan C – 1903015

Vivek Vishal Sharma - 19030150

## Business Context – Brief background

AST is the insurance company, protects people from life’s uncertainties with more than 113 million proprietary policies. In insurance policy shopping, customer will receive number of quotes with different coverage options before purchasing a plan. Every time a customer asks for a quote with different product combinations and purchase decision can be a lengthier process. Based on past history customer shopping history with multiple quotes and customer characteristics, AST would like to predict the purchase decision of customer. If the purchase prediction is sooner in the shopping window, the quotation cycle is shortened and AST is less likely to lose the customer’s business. The prediction problem considered in this project is AST’s car insurance product.

## Business Objective

Due to the rapid growth in automobiles, vehicle insurance has emerged as a critical target for insurance companies. It is crucial to improve business performance by acquiring new customers, retaining new customers and faster response in finalizing the quote. Minimizing the conversion of existing customers to other insurance companies is an important factor which increases the company profit. So, it is very important to increase the business efficiency by improving quality of service with quick response time with right product options according to customer characteristics. With a probabilistic prediction of the purchase, a firm can direct its marketing and sales resources to the most promising target set.

The data available for testing was identical to the training set, except it was a different set of customers, the purchase point was excluded (since it was the value to be predicted),

and an unknown number of quotes were also excluded.

## Analytics Objectives

The analytics objective of the project is to predict the purchase point (Yes / No) by using supervised learning method of machine learning techniques. Supervised learning maps an input to an output based on example input-output pairs. It infers a function from labelled training data consisting of a set of training examples (number of records). The data available for training a model is the history of car insurance quotes that each customer reviewed before making a purchase, the options they actually purchased (the "record\_type"), and data about the customer and their car. The data contains various demographic and situational factors. The demographic factors include the factors like car and home ownership, general risk factor, age of eldest and youngest member of the buyer group etc. The situation factors are the access location, day and time of soliciting quote etc. The product factors are the product features/attributes being queried for and of course, the price.

## Specific questions that you seek to answer using analytics techniques (Please check if the questions can be answered using standard cross tabs. If so, you need to re-evaluate your questions)

* Identify and classify the target customers who will be interested to buy the insurance based on quote history and customer characteristics.
* Influence the prospective customer with additional product options to give additional value to customer and thus cross-selling of product options.
* The most important part of this analysis is getting the answers to the question will this customer going to buy the insurance now? There are multiple factors that impact the decision. Few of these are categorical and others are quantitative or ordinal. In all, there are 25 identified variables that will help us to predict the outcome.

## Overview of the data set – source, no. of records, fields etc.

Source: <https://www.kaggle.com/akhilups/insurance-product-purchase-prediction>

No. of records: 665,249

Variable Fields / Columns: 25

Note: The source contains individual file for training and testing. In the project, we have used train data file itself for testing with k-fold method

List of fields/attributes in the train data set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl. No. | Variable Name | Definition | Scale | Python data type |
| 1 | customer\_ID | A unique identifier for the customer | Ordinal | Numeric |
| 2 | shopping\_pt | Unique identifier for the shopping point of a given customer (1 – 13) | Ordinal | Numeric |
| 3 | record\_type | Purchase point (0=quote, 1=purchase) | Nominal | Category |
| 4 | day | Day of the week (0-6, 0=Monday) | Ordinal | Numeric |
| 5 | time | Time of day (HH:MM) | DateTime | DateTime |
| 6 | state | Two letter state code where shopping occurred | Nominal | Category |
| 7 | location | Location ID where shopping occurred | Ordinal | Numeric |
| 8 | group\_size | How many people will be covered under the policy (1, 2, 3 or 4) | Nominal | Category |
| 9 | homeowner | Whether the customer owns a home or not (0=No, 1=Yes) | Nominal | Category |
| 10 | car\_age | Age of the customer’s car (in years) | Ratio | Numeric |
| 11 | car\_value | How valuable was the customer’s car when new; classified into 9 groups (a – i) | Nominal | Category |
| 12 | risk\_factor | An ordinal assessment of how risky the customer is (1, 2, 3, 4) | Nominal | Category |
| 13 | age\_oldest | Age of the oldest person in customer's group in years | Ratio | Numeric |
| 14 | age\_youngest | Age of the youngest person in customer’s group in years | Ratio | Numeric |
| 15 | married\_couple | Does the customer group contain a married couple (0=no, 1=yes) | Nominal | Category |
| 16 | C\_previous | What the customer formerly had or currently has for product option C (0=nothing, 1, 2, 3,4) | Nominal | Category |
| 17 | duration\_previous | how long (in years) the customer was covered by their previous issuer | Ordinal | Numeric |
| 18 | A | the coverage options A; (0-1-2) | Nominal | Category |
| 19 | B | the coverage options B; (0-1) | Nominal | Category |
| 20 | C | the coverage options C; (1-2-3-4) | Nominal | Category |
| 21 | D | the coverage options D; (1-2-3) | Nominal | Category |
| 22 | E | the coverage options E; (0-1) | Nominal | Category |
| 23 | F | the coverage options F; (0-1-2-3) | Nominal | Category |
| 24 | G | the coverage options G; (1-2-3-4) | Nominal | Category |
| 25 | cost | cost of the quoted coverage options | Ratio | Numeric |

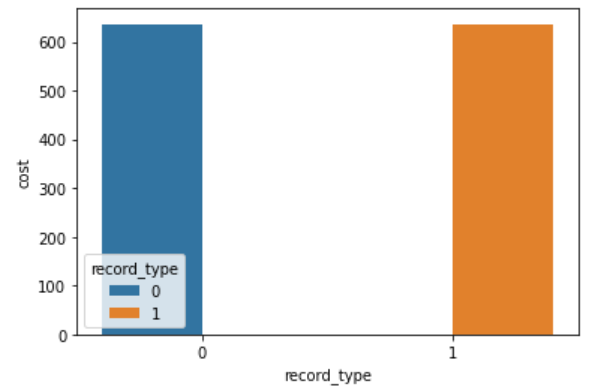
## Explanation of your data set (use frequency distributions and / or cross tabs)

The data contains lot of missing values and null values. Also the data types are not right as per the values. The raw data type details with missing values are tabled below

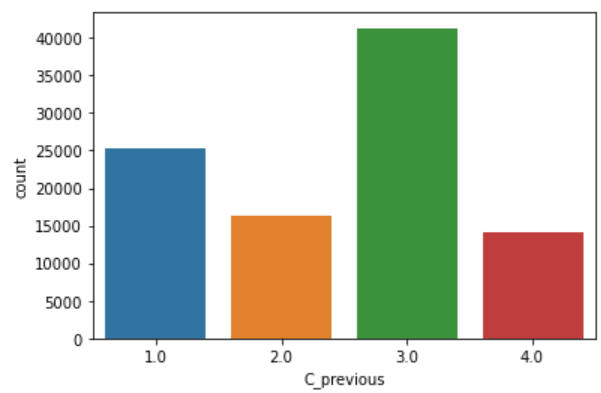
|  |  |
| --- | --- |
| **Data Types** | **Missing Values** |
| <class 'pandas.core.frame.DataFrame'>  RangeIndex: 665249 entries, 0 to 665248  Data columns (total 25 columns):  # Column Non-Null Count Dtype  --- ------ -------------- -----  0 customer\_ID 665249 non-null int64  1 shopping\_pt 665249 non-null int64  2 record\_type 665249 non-null int64  3 day 665249 non-null int64  4 time 665249 non-null object  5 state 665249 non-null object  6 location 665249 non-null int64  7 group\_size 665249 non-null int64  8 homeowner 665249 non-null int64  9 car\_age 665249 non-null int64  10 car\_value 663718 non-null object  11 risk\_factor 424831 non-null float64  12 age\_oldest 665249 non-null int64  13 age\_youngest 665249 non-null int64  14 married\_couple 665249 non-null int64  15 C\_previous 646538 non-null float64  16 duration\_previous 646538 non-null float64  17 A 665249 non-null int64  18 B 665249 non-null int64  19 C 665249 non-null int64  20 D 665249 non-null int64  21 E 665249 non-null int64  22 F 665249 non-null int64  23 G 665249 non-null int64  24 cost 665249 non-null int64  dtypes: float64(3), int64(19), object(3) | 0  0  0  0  0  0  0  0  0  0  1531  240418  0  0  0  18711  18711  0  0  0  0  0  0  0  0 |

Data has been cleaned with iterative approach by imputing values for missing values, dropping few unwanted variables and converting the data types.

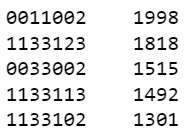
From the initial uni-variate exploratory analysis, we understood that cost is not the prime factor for the decision as the purchasing decision of 0 or 1 distributed for the all the cost range

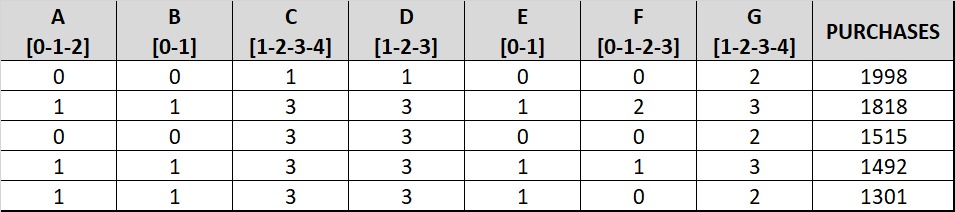


We have created new data frame ‘purchase’ by extracting only purchased policy option (record\_type = 1). By using this data\_frame, analysed the previous ‘C’ product option. It is found that option 3 is mostly used in C product option.

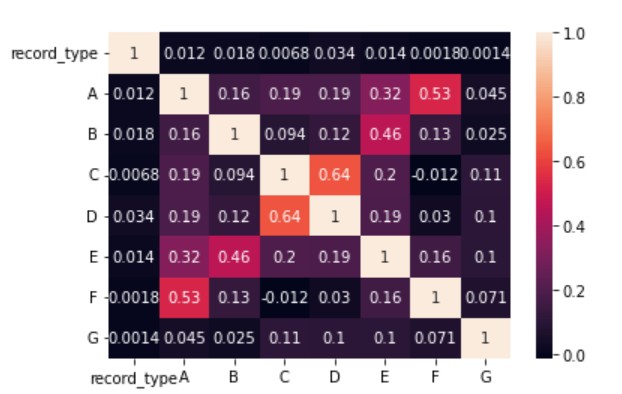


Also, product option is uniquely identified and counted. The top are





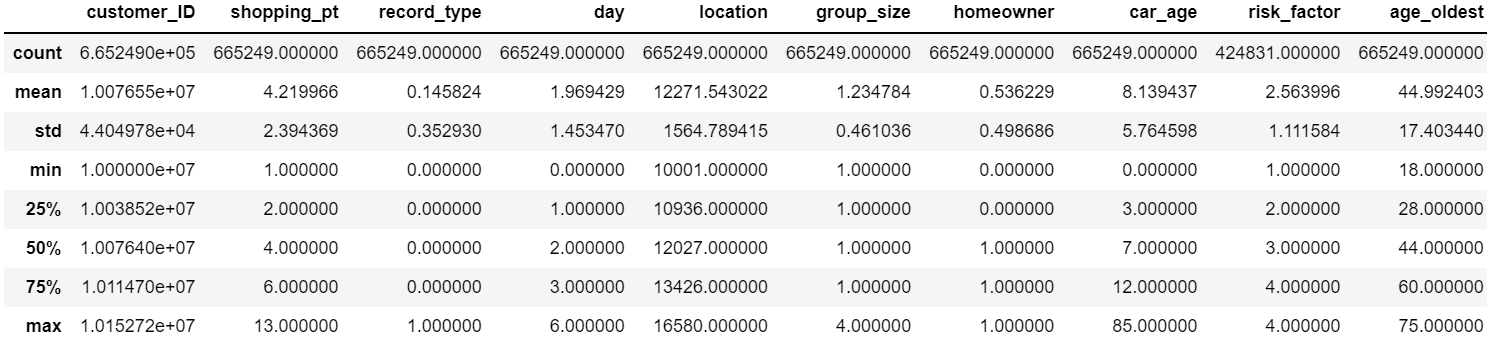
To understand the product options correlations with one another and with purchasing point is explored with heat map. From the analysis product option ‘C’ and ‘D’ are highly correlated. It is also evident from the above top 5 product option.

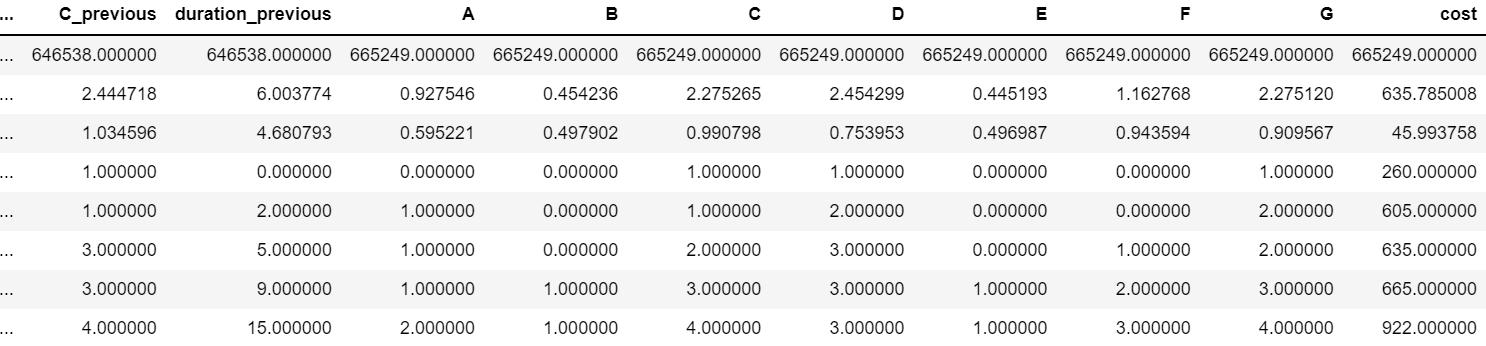


## What attributes do you plan to use? Have you checked on options of deriving new variables?

All the attributes of 25 variables are used. For exploratory analysis, we have derived new variable called ‘CombOpt’ to combine the production options ‘A’, ‘B’, ‘C’, ‘D’, ‘E’, ‘F’, ‘G’ to analyse the top 5 product options purchased.

## How is each attribute distributed? Are there too many different values? Too many same values? Too many null values? What’s your plan to deal with these issues?





Only the Car Age has values distributed with 84 values and the mean is 8 years only. Maximum number of years represent the Vintage cars.

Null values are present in car\_value, risk\_factor, C\_previous, duration\_previous as mentioned in section 6.

* car\_value is converted into category and blank is one of the category code assigned
* As 1/4th of data is null value for risk\_factor, that column is dropped
* Other null values are imputed with the strategies ‘constant’ and ‘median’

## Are any of the attributes correlated highly? How will you ensure that the correlation does not adversely affect your results?

From the heat map depicted in section 6, the product options are highly correlated which does not have any adverse effect. In fact the high correlation is expected for cross-selling

## If you plan to use Prediction, what is your target variable? Check the distribution. If it is skewed towards one outcome, how do you plan to address this in your plan?

Target variable is the purchase point (variable name: record\_type) which is the categorical variable to be predicted based on customer characteristics during every shopping point (quote/interactions)

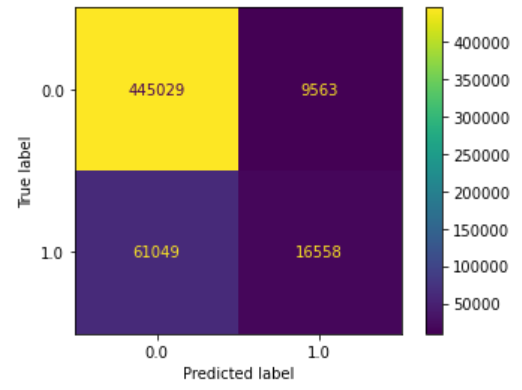
This is represented by dependent variable ‘record\_type’ {0, 1}

## What is your analytical methodology? Which techniques do you plan to use? Why?

We are using the three prediction models viz. Decision Tree, Logistic Regression and Neural Network. Apart from the prediction, we have also used the Market Basket Analysis to influence the product options. The prediction would further be evaluated using ROC curve and confusion matrix.

## Decision Tree

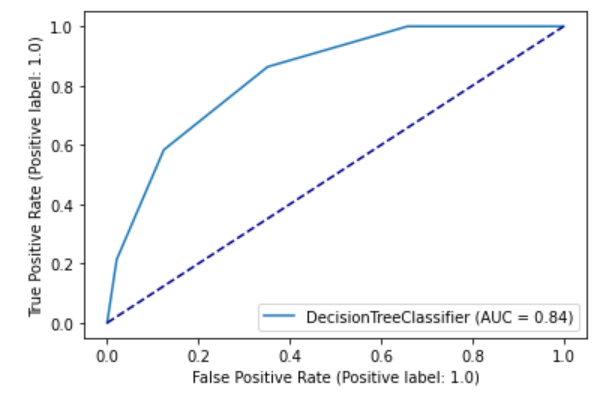
The confusion matrix from the training is below



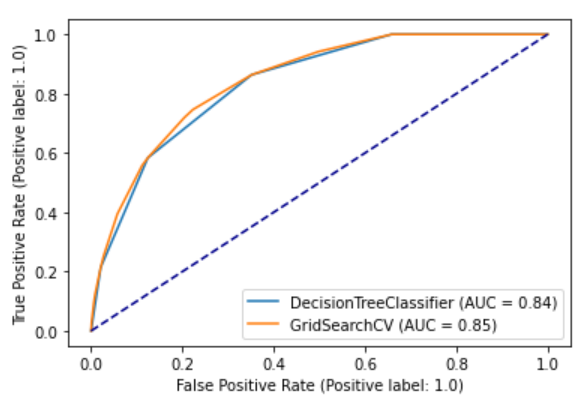
**Accuracy:**

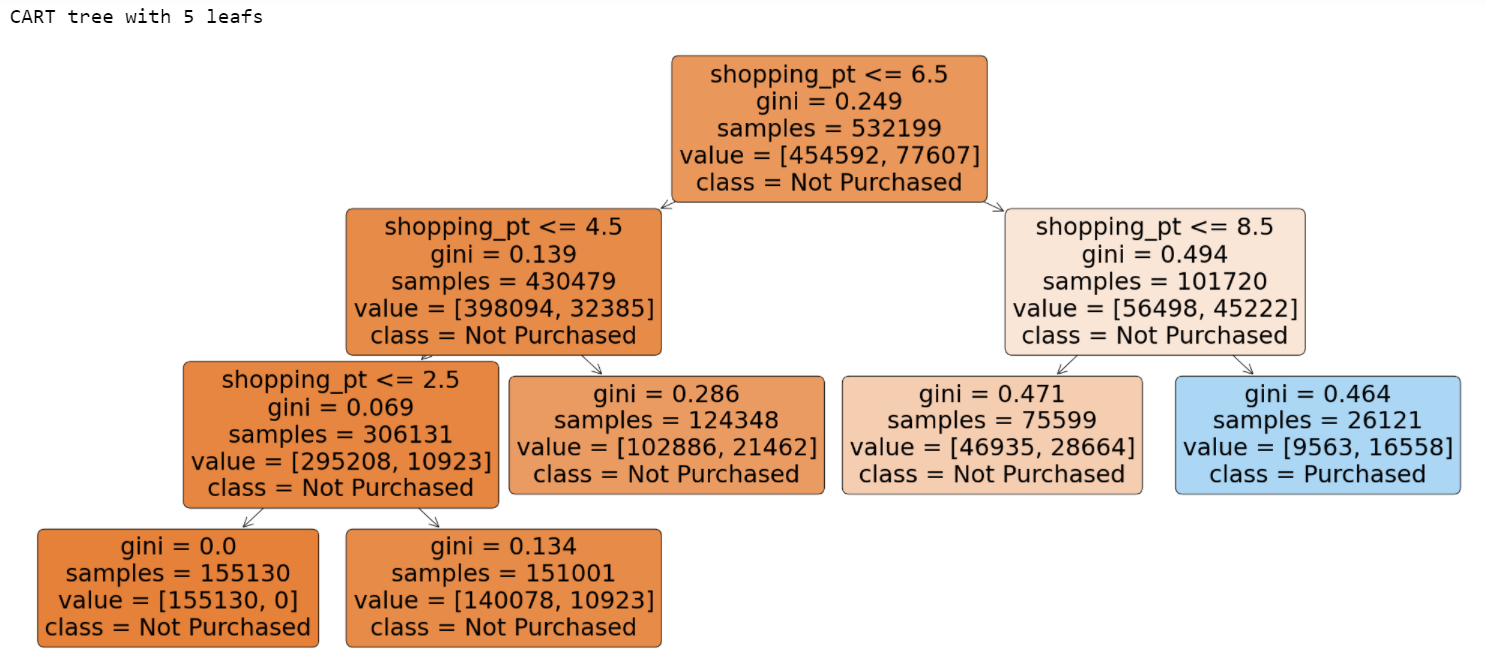
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Error Rate | Sensitivity | Specificity | Precision |
| Decision Tree | 0.8673 | 0.1327 | 0.2134 | 0.979 | 0.6339 |

**ROC Curve**



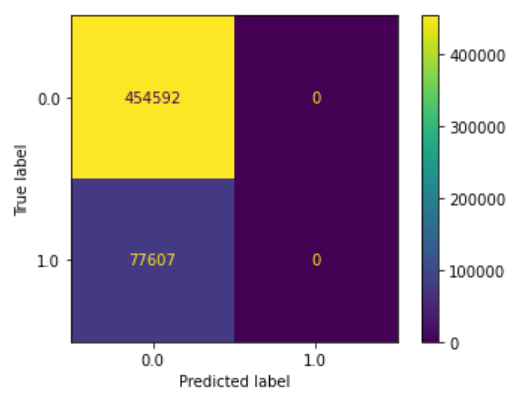
**ROC Curve with Grid Search**





## Logistic Regression

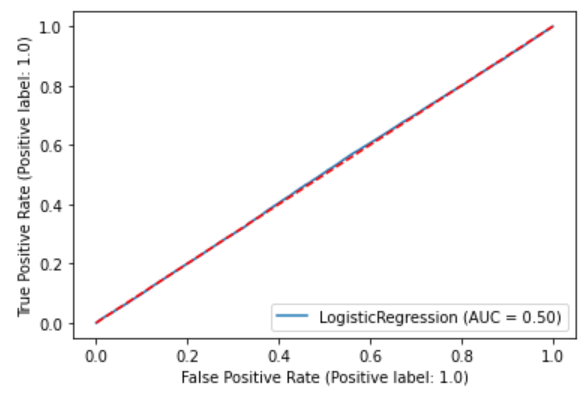
The confusion matrix from the training is below



**Accuracy:**

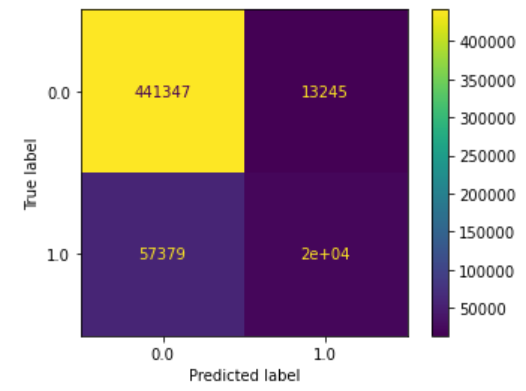
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Error Rate | Sensitivity | Specificity | Precision |
| Logistic Regression | 0.8542 | 0.1458 | 0.0 | 1.0 | undefined |

**ROC Curve**



## Neural Network

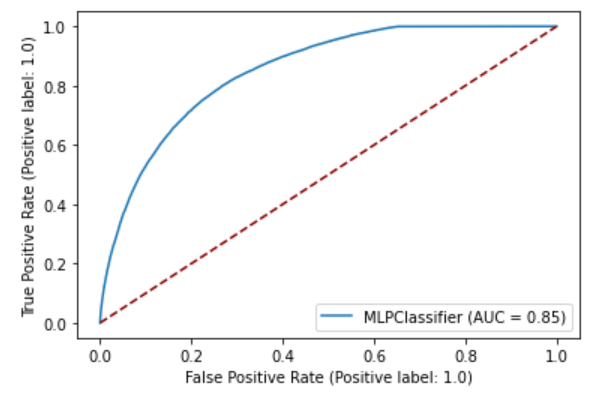
The confusion matrix from the training is below



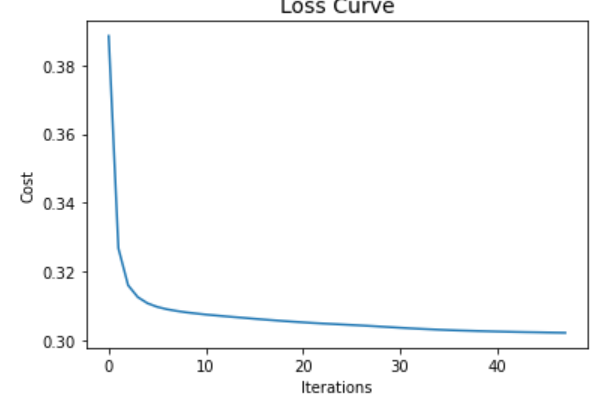
**Accuracy:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Error Rate | Sensitivity | Specificity | Precision |
| Neural Network | 0.8674 | 0.1326 | 0.2612 | 0.9709 | 0.6049 |

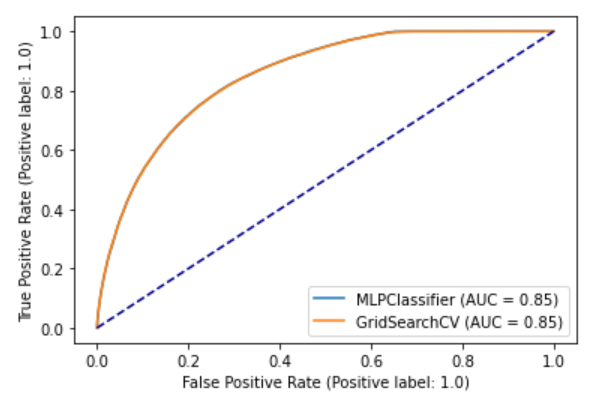
**ROC Curve**

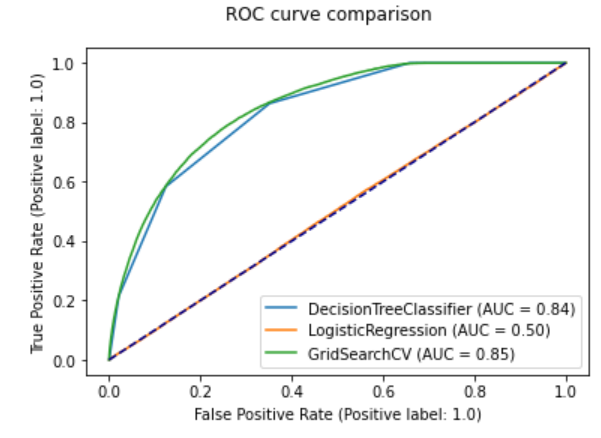


**Loss Curve:**



## ROC Curve comparison and model evaluation



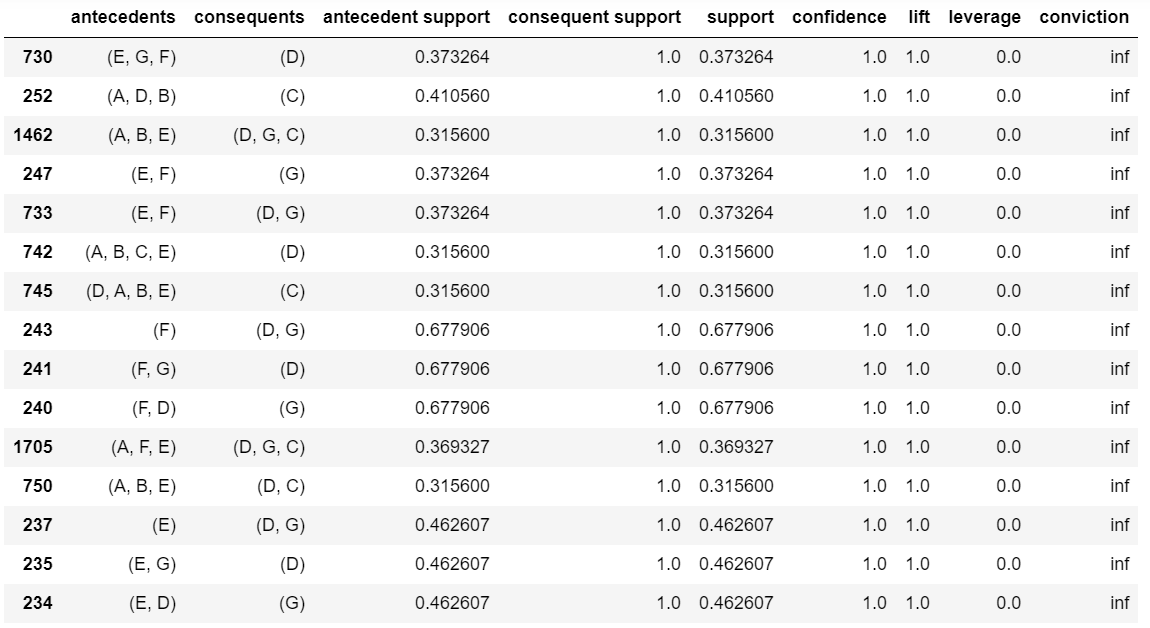


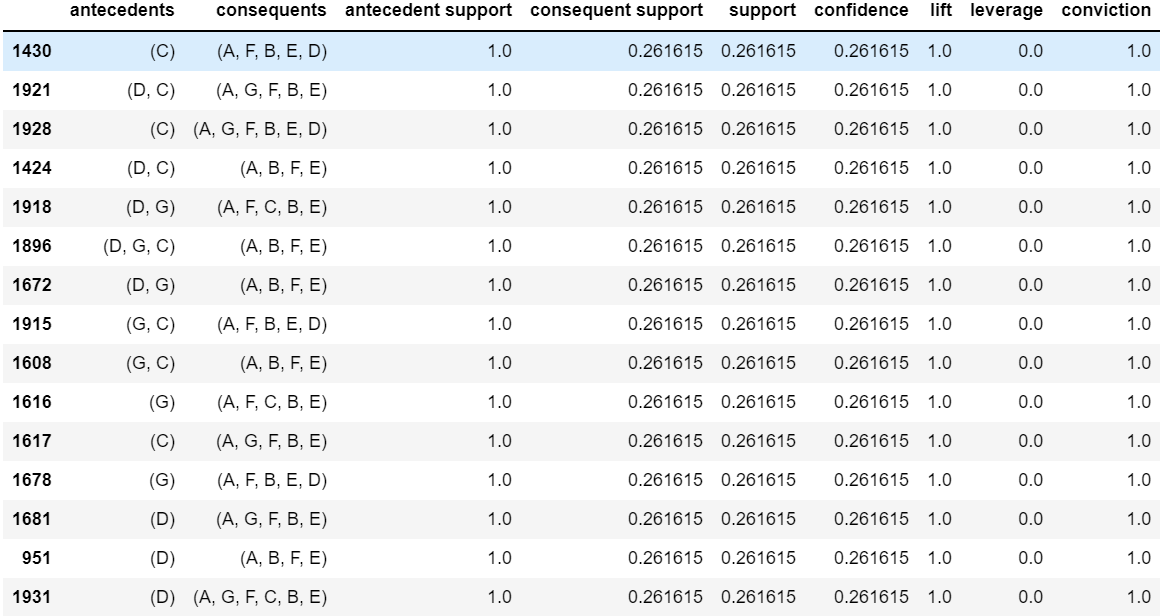
All the above models are evaluated and based on ROC Curve and confusion matrix evaluation, Decision Tree and Neural Network model have high ‘Area Under Curve’ (AUC). Generally Logistic Regression model is best suited for predicting categorical variable. However, based on the data provided, we recommend the AST insurance company to use the Decision tree / Neural Network model to evaluate the purchase prediction

## Market Basket Analysis

Market Basket Analysis (MBA) was done to influence the related product option while purchasing the policy.

"When 'C', 'D' and 'G' are Consequent, Consequent\_Support is 1. This implies that 'C', 'D' and 'G' are always bought"





## How will results from your analytics plan help you solve your Business Problem?

This project will help the AST insurance company to predict the purchase to shorten the number of quotes and also cross sell multiple product options in the same policy.