Siddhi Padekar

25.07.2021

InSURANCE FIRM

Data Analysis

Table of Contents Contents Executive

|  |  |
| --- | --- |
| Summary |  |
| Introduction | 2 |
| Data Description . | 2 |
| Sample of the dataset: | 2 |
| Exploratory Data Analysis | 3 |
| Types of variables in the data frame | 3 |
| Missing values in the dataset | 3 |
| Univariate Analysis , Bivariate analysis, Multivariate Analysis | 4 |
| Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network | 7 |
| CART Analysis | 7 |
| Random Forest Analysis | 8 |
| Artificial Neural Network | 9 |
| Compare all the models and write an inference which model is best/optimized. | 10 |
| Inference: Based on the whole Analysis, what are the business insights and recommendations | 10 |

Executive Summary

An Insurance firm deals with different agencies and product plans for its customers in America, Asia and Europe. The dataset consists of how customers avail claims of different products of firm. In this problem statement, we will study how to predict the claim status by using different performance models.

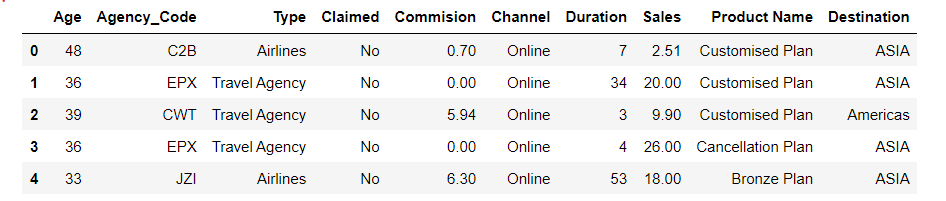
Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset, predict claim status using CART, RF & ANN, and compare the models' performance. The data consists of 5 different Insurance product plan and their sales value, durations commissions of different agencies in different countries.

Data Description

1. Target: Claim Status (Claimed)  
   2. Code of tour firm (Agency\_Code)  
   3. Type of tour insurance firms (Type)  
   4. Distribution channel of tour insurance agencies (Channel)  
   5. Name of the tour insurance products (Product)  
   6. Duration of the tour (Duration)  
   7. Destination of the tour (Destination)  
   8. Amount of sales of tour insurance policies (Sales)  
   9. The commission received for tour insurance firm (Commission)  
   10. Age of insured (Age)

Sample of the dataset:



Dataset consists of 10 columns

Exploratory Data Analysis

Types of variables in the data frame

Data columns (total 10 columns):

Age int64

Agency\_Code object

Type object

Claimed object

Commision float64

Channel object

Duration int64

Sales float64

Product Name object

Destination object

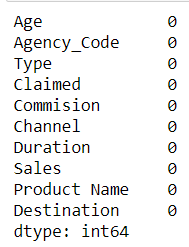
dtypes: float64(2), int64(2), object(6)

There are total 3000 rows and 10 columns in the dataset. Out of 10, 6 columns are of object type ,2 columns are integers and 2 are float data type.

Check for missing values in the dataset:

RangeIndex: 3000 entries, 0 to 2999

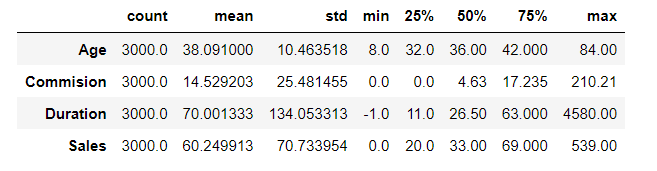
Data columns (total 10 columns):



dtypes: float64(2), int64(2), object(6)

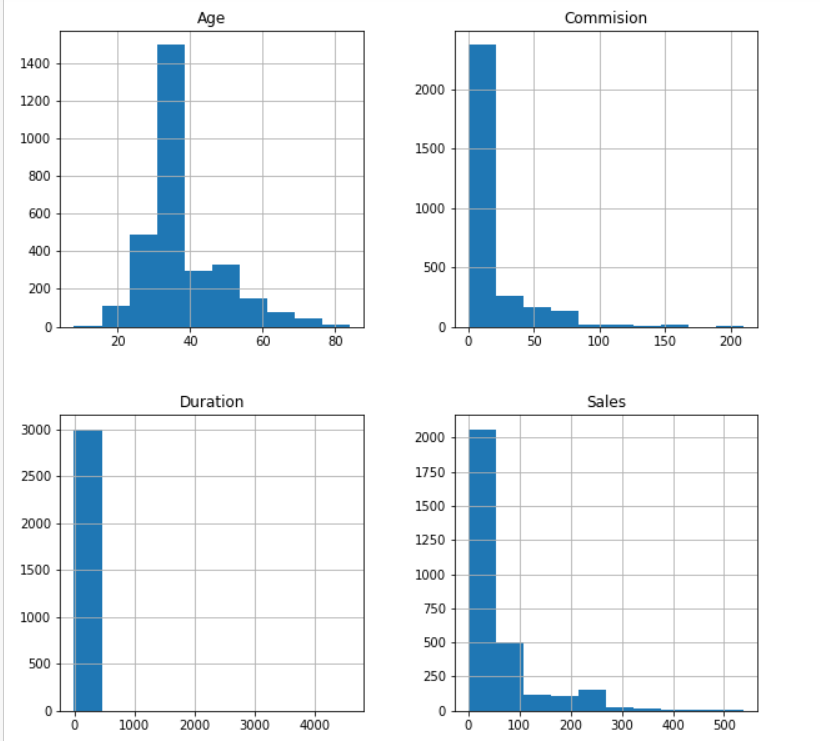
* No missing value present in dataset

Mean, Standard deviation , Min, Max



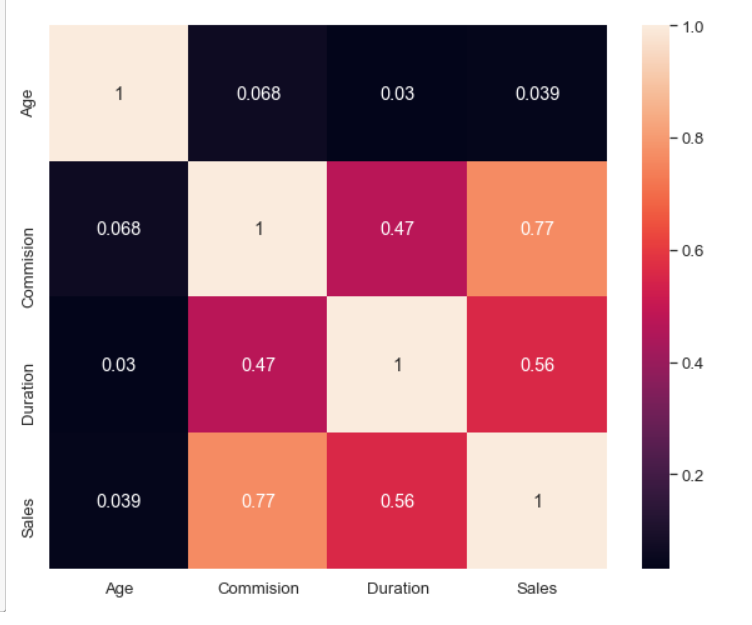
Univariate Analysis

Histograms



Multivariate analysis

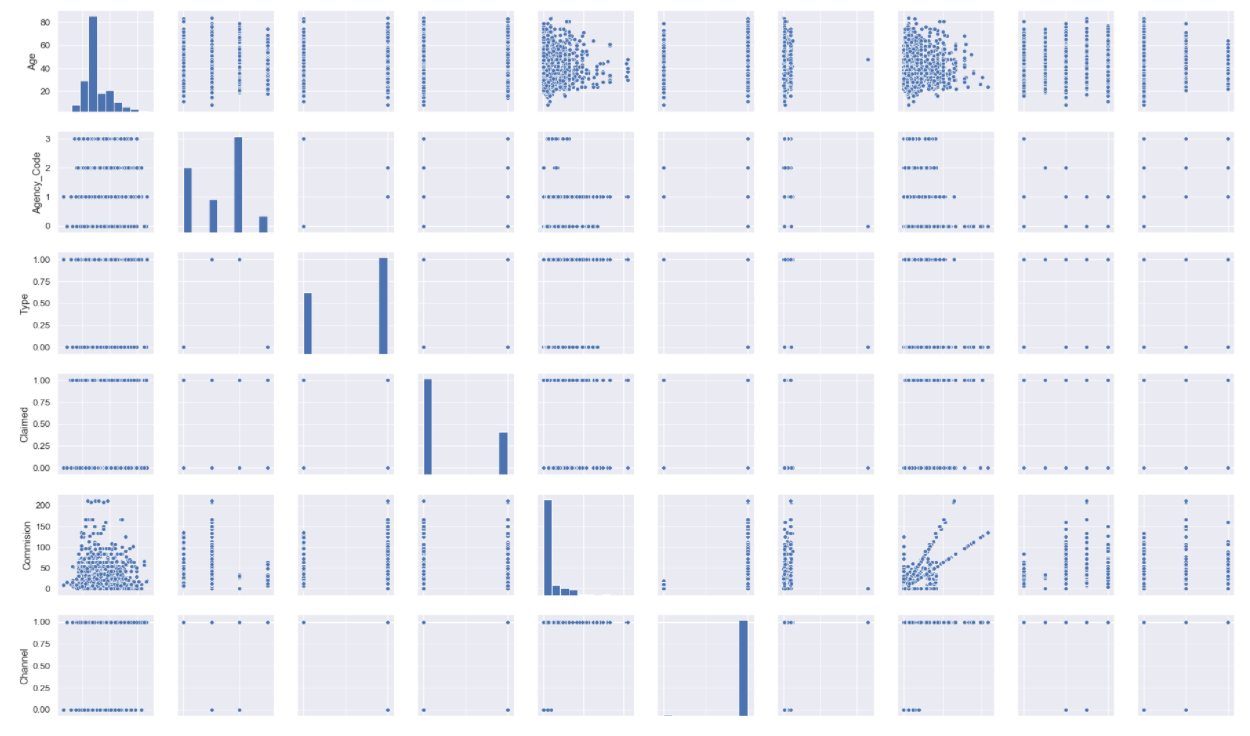
Correlation Plot



From the correlation plot, we can see that various attributes of the claims are highly correlated to each other. Correlation values near to 1 are highly positively correlated. Correlation values near to 0 are not correlated to each other.

Pair Plot

Pairplot visualizes given data to find the relationship between them where the variables can be continuous or categorical.



Data Split-

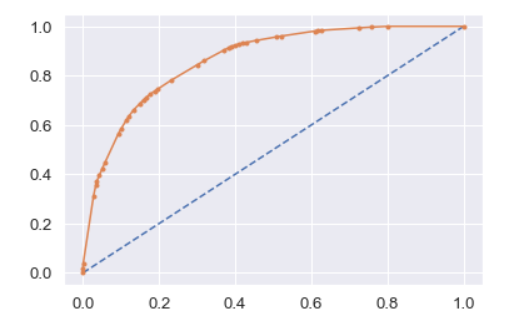
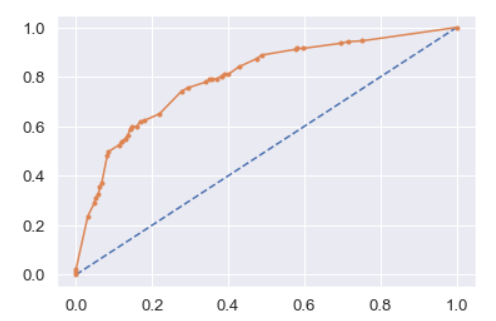
Data is split into training and testing using claimed attribute to predict claim status using different models

CART analysis

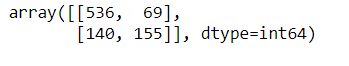
AUC score

|  |  |
| --- | --- |
| Train data (Development data) | Test data (Hold out data) |
| 86% | 79% |

ROC Curve

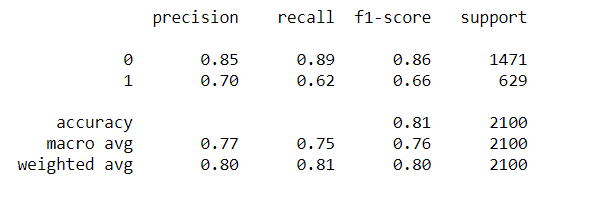
 

Confusion matrix

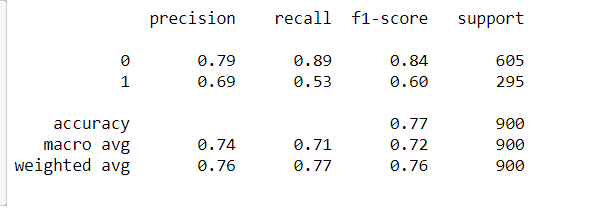
 

Classification model

Train data



Test Data



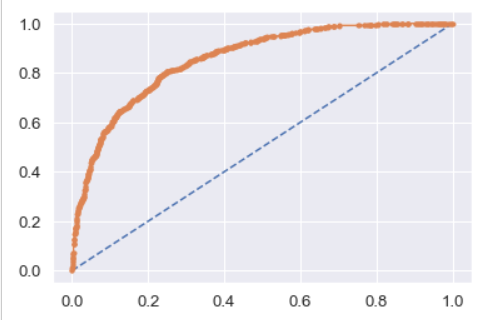
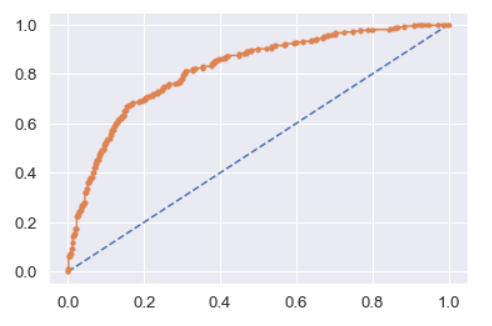
CART model shows train data is AUC is more than test data. Also if we observe the area coverage of train data in the graph it is more than the test data. From classification model we can also see precision score and f1score of train data is more than precision score and f1score of test data.

Random Forest Analysis

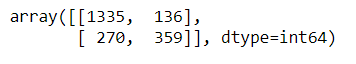
AUC score

|  |  |
| --- | --- |
| Train data (Development data) | Test data (Hold out data) |
| 86% | 82% |

ROC Curve

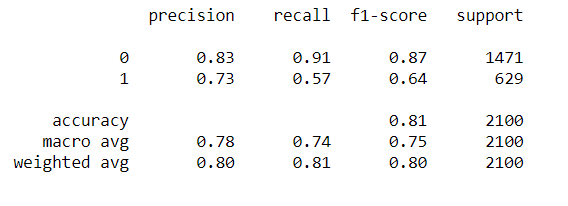
 

Confusion matrix

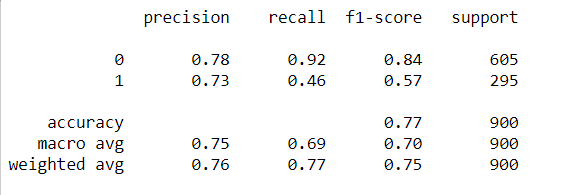
 

Classification Matrix

Train Data

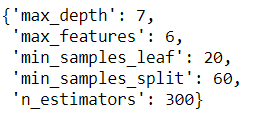


Test Data



Random Forest Analysis model shows train data AUC is more than test data. Also if we observe the ROC curve area coverage of train data in the graph it is more than the test data. From classification model we can also see precision score and f1score of train data is more than precision score and f1score of test data.

Also best grid parameters are



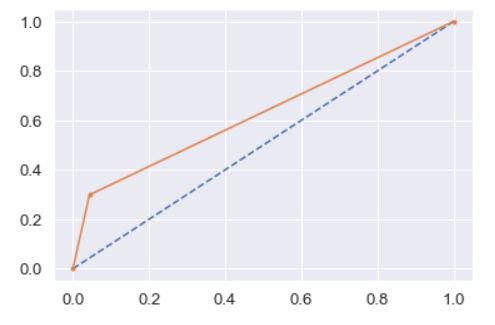
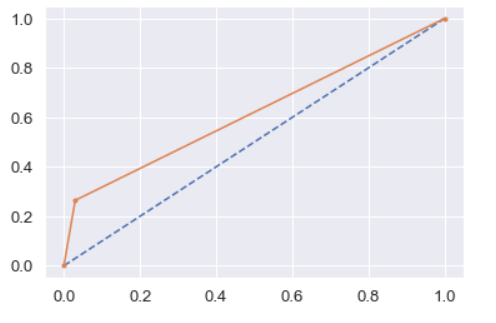
testing data is underperforming than the training data

Artificial Neural Network

AUC score

|  |  |
| --- | --- |
| Train data (Development data) | Test data (Hold out data) |
| 80% | 79% |

ROC Curve

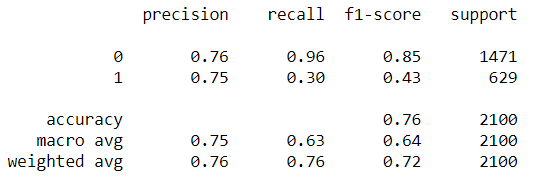
 

Confusion matrix

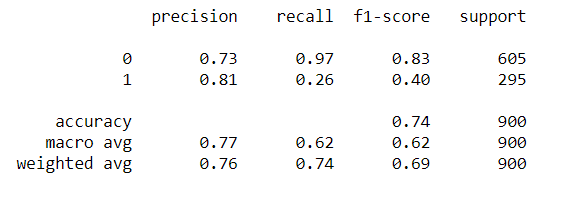
 

Classification Matrix

Train Data



Test Data



Artificial Neural Network model shows train data AUC are almost equal to test data. Also if we observe the ROC curve area coverage of train data in the graph are almost equal to test data. From classification model we can also see precision score and f1 score of train data is more than precision score and f1score of test data.

* If we compare all the Analysis i.e CART analysis , Random forest analysis and Artificial Neural Network we can see that test data is under performing than the test data.

Trained data scores

|  |  |  |  |
| --- | --- | --- | --- |
|  | CART | Random Forest | Artificial Neural Network |
| AUC | 86% | 86% | 80% |
| Precision score | 0.85 | 0.83 | 0.76 |
| F1 score | 0.86 | 0.87 | 0.85 |
| Accouracy | 81% | 81% | 76% |

* Cart model and Random forest model gives almost same accuracy level. The best model in hand can be CART if we see the precision score.