



Capstone Project

Predicting Best Insurance plan

MACHINE LEARNING FOUNDATION TRAINING (BATCH 04)

SHASHEN WIJESINGHE-REG.NO-203

Table of Contents

Chapter 01- Introduction-----	02
Chapter 02-Data-----	03
Chapter 03-Methodology-----	05
Chapter 04-Results-----	07
Chapter 05-Discussion and Conclusion-----	11

Table of Figures

Figure 1:Newly Purchased Insurance Plans in IT-Sector for Mar-Aug 2022.....	2
Figure 2:Correlation of the variables	4
Figure 3:Structure of Neural Network	6
Figure 4:Confusion Matrix for Logistic Regression	7
Figure 5:Classification Report for Logistic Regression	7
Figure 6:Confusion Matrix for Support Vector Machines.....	8
Figure 7:Classification Report for Support Vector Machines	8
Figure 8:Training the NN.....	9
Figure 9:Classification Report for NN.....	9
Figure 10:Confusion Matrix for Decision Tree Classifier	10
Figure 11:Classification Report for Decision Tree Classifier.....	10

Table of Tables

Table 1:Data fields for the dataset, description and datatype	3
Table 2:Data Describe of Non-Categorical Variables	3
Table 3:Package Defining Criteria.....	4
Table 4:Customers included in each PLAN.....	4
Table 5:Total Plans	5
Table 6: Inputs and Output for the ML Model	5

INTRODUCTION

Climax Insurance(pvt) Ltd is a leading insurance company in Sri Lanka for more than 20 years. Most of the customers are from western province. As a result of newly increased package price plans for their customers, they have identified some lack of interest to buy new insurance plans. The management need to identify this problem and predict the most suitable plan for their customers and promote them.

Ravindu, who is working as a business analyst for more than 5 years is requested to make this decision as soon as possible and he is going to use his machine learning knowledge which gained from the recently conducted and finished machine learning foundation training.

As per final meeting with the management, company confirmed that this lack of interest is only from the IT-sector in Kaluthara district only.

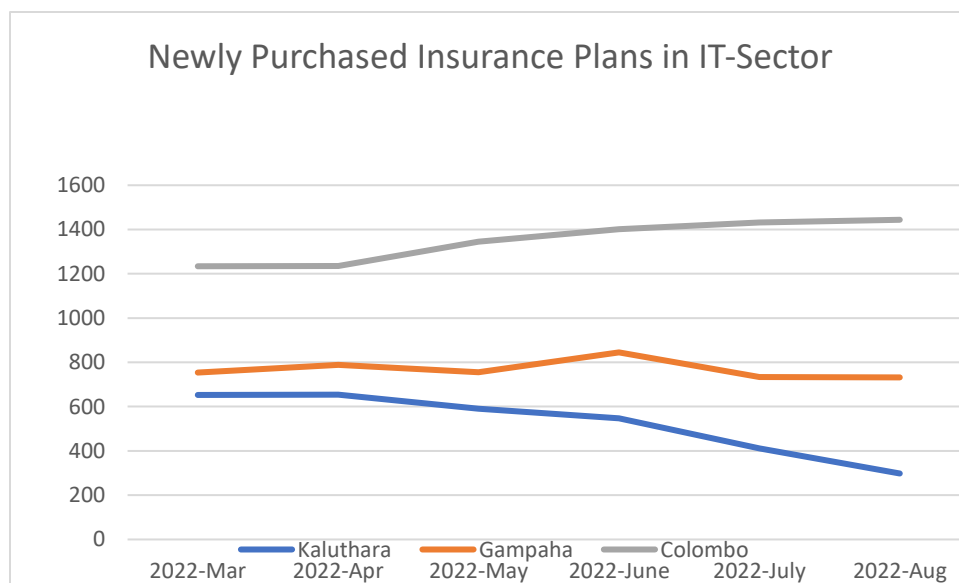


Figure 1: Newly Purchased Insurance Plans in IT-Sector for Mar-Aug 2022

DATA

Since the issue is with IT-sector in Kaluthara, Ravindu has focused only the data from Kaluthara district, and the all these customers are in IT-sector. From the initially filled form, when requesting for an insurance plan Ravindu has extracted data with following data fields.

Table 1:Data fields for the dataset, description and datatype

Data Field	Description	Data Type
Age	Age of the Customer	Int64
Sex	Gender of the Customer	Object
bmi	Body Mass Index calculated from the wight and height of the employee	Float64
children	No. of children for the Customer	Int64
smoker	Whether customer is smoking or not	Object
region	Region in the district	Object
charges	Requested plan by the Customer	Float64

Total dataset contains 1338 records and 7 columns. There are no any null data in this dataset.

Below describes the data of non-categorical variables in the dataset.

Table 2:Data Describe of Non-Categorical Variables

index	age	bmi	children	charges
count	1338	1338	1338	1338
mean	39.20703	30.6634	1.094918	13270.42227
std	14.04996	6.098187	1.205493	12110.01124
min	18	15.96	0	1121.8739
25%	27	26.29625	0	4740.28715
50%	39	30.4	1	9382.033
75%	51	34.69375	2	16639.91252
max	64	53.13	5	63770.42801

By considering the package price variation, PLAN is introduced, and new column created in the dataset. Below are the criteria of define the PLAN.

Table 3:Package Defining Criteria

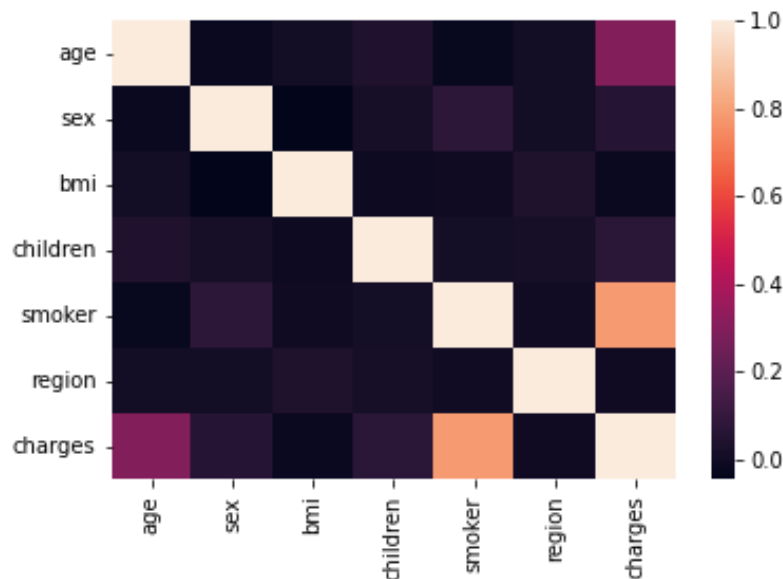
Price or Price Range	PLAN
Below 5000	PKG1
Between 5000 and 10000	PKG2
Between 10000 and 15000	PKG3
More than 150000	PKG4

After defining the PLAN, below is the customer interest on various PLAN types.

Table 4:Customers included in each PLAN

PLAN	No. of Customers
PKG1	359
PKG2	358
PKG3	353
PKG4	268

Now need to identify the correlation of each variable to the purchased plan.



According to above figure correlation of bmi variable to charges variable is -0.02 and therefore bmi variable is dropped.

METHODOLOGY

1)Data Gathering

1338 records collected from the filled forms with 7 columns to create the dataset. As per the correlation bmi is dropped and there are only 7 columns. No null values in the dataset.

2)Data Preprocessing

Since there is no any null values, no need to consider the replacement for the rows. The variable ‘bmi’ can be dropped since that will not produce much correlation to the charge variable. From the PLAN variable, below is the output of the unique value counts.

Table 5:Total Plans

Plan Type	Total Plans	Encoded Value
PKG1	359	0
PKG4	358	3
PKG2	353	1
PKG3	268	2

Since classes are almost equally distributed, no need to resample the dataset to avoid class imbalances in the dataset. Both categorical and non-categorical data in the dataset, therefore we need to encode the labels of non-categorical data. According to the Table 2, age variable has a considerable data range. Therefore, age column has scaled using MinMaxScaler.

After preprocessing data, following are the inputs (X Variables) and output(Y Variable) for the ML model.

Table 6: Inputs and Output for the ML Model

X Variables	Y Variable
age	PLAN
sex	
children	
smoker	
region	

2)Solution Approach and Used Tools

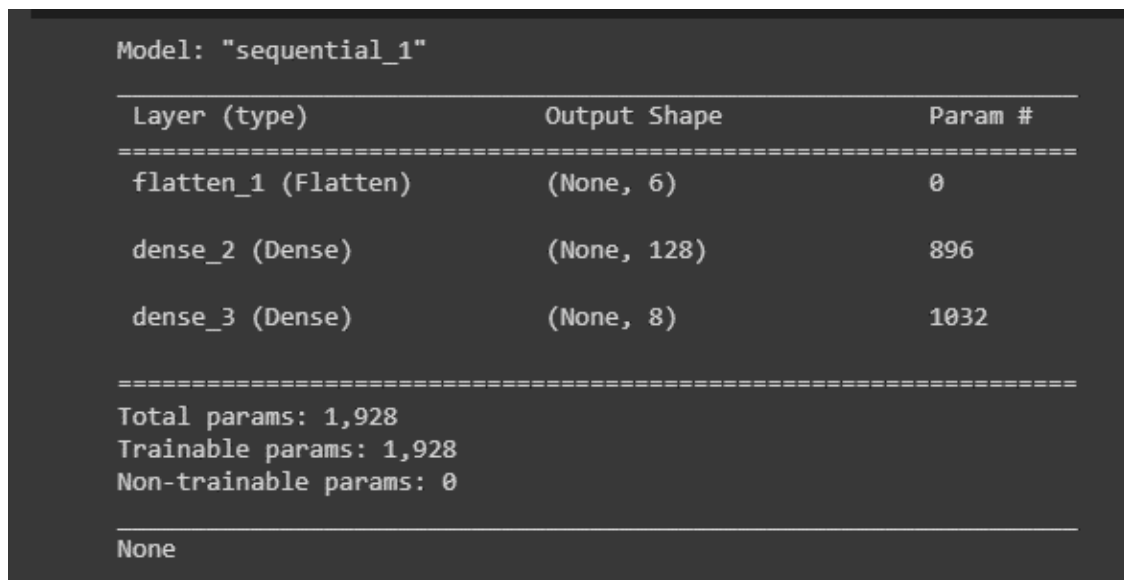
After preprocessing the data, now all set to develop the model. In here three machine learning techniques have used. Following are the used ML algorithms with their libraries.

1. Logistic Regression- 'lbfgs' solver has used, and max iterations are set to 1000
2. Support Vector Machines
3. Neural Network
4. Decision Tree Classifier-max depth considered as 10 and minimum sample split is 2.

For all above algorithms, the dataset has split to train and test. 30% of total dataset considered as test dataset to verify the model and rest 70% is train dataset and it has used to train the model.

- Neural Network (NN)

To develop the neural network three layers have defined. Below Figure 3 depicts the structure of the Neural Network.



```
Model: "sequential_1"
Layer (type)                Output Shape              Param #
=====
flatten_1 (Flatten)         (None, 6)                 0
dense_2 (Dense)              (None, 128)              896
dense_3 (Dense)              (None, 8)                1032
=====
Total params: 1,928
Trainable params: 1,928
Non-trainable params: 0
None
```

Figure 3:Structure of Neural Network

Further batch_size has used as 128 and 50 epochs are trained when compiling the NN.

RESULTS

1. Logistic Regression

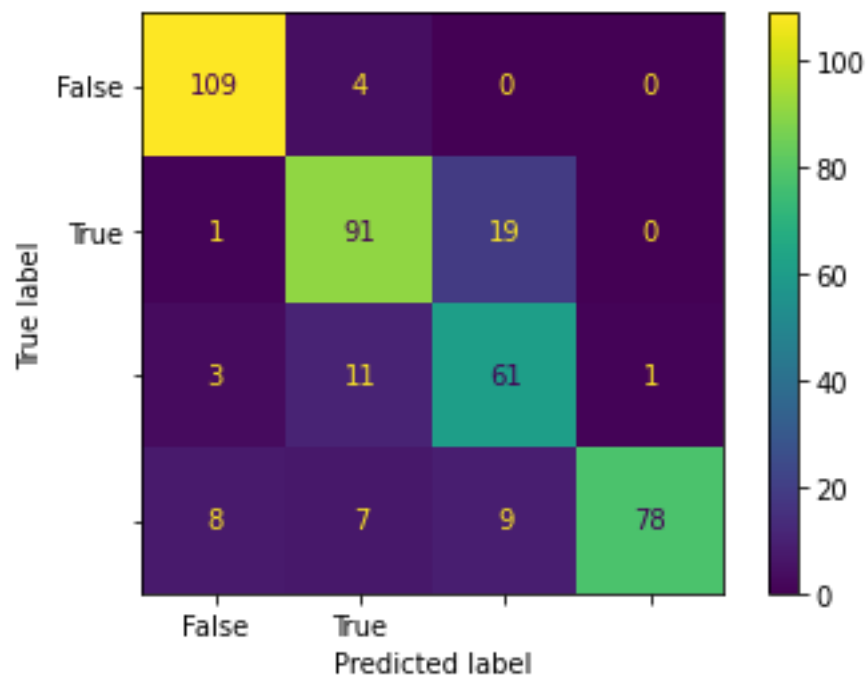


Figure 4:Confusion Matrix for Logistic Regression

	precision	recall	f1-score	support
0	0.91	0.94	0.92	113
1	0.82	0.86	0.84	111
2	0.75	0.84	0.80	76
3	0.93	0.76	0.84	102
accuracy			0.85	402
macro avg	0.85	0.85	0.85	402
weighted avg	0.86	0.85	0.85	402

Figure 5:Classification Report for Logistic Regression

2. Support Vector Machines

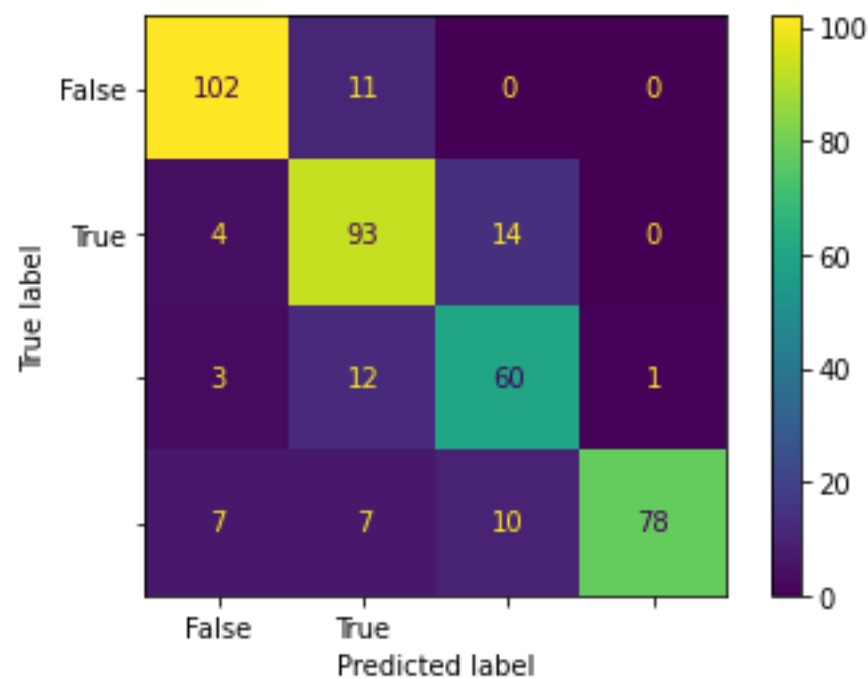


Figure 6:Confusion Matrix for Support Vector Machines

	precision	recall	f1-score	support
0	0.88	0.90	0.89	113
1	0.76	0.84	0.79	111
2	0.71	0.79	0.75	76
3	0.99	0.76	0.86	102
accuracy			0.83	402
macro avg	0.83	0.82	0.82	402
weighted avg	0.84	0.83	0.83	402

Figure 7:Classification Report for Support Vector Machines

3. Neural Network (NN)

```
Epoch 1/50
8/8 [=====] - 1s 36ms/step - loss: 1.4312 - sparse_categorical_accuracy: 0.4380 - val_loss: 1.0499 - val_sparse_categorical_accuracy: 0
Epoch 2/50
8/8 [=====] - 0s 8ms/step - loss: 0.9719 - sparse_categorical_accuracy: 0.6688 - val_loss: 0.8442 - val_sparse_categorical_accuracy: 0
Epoch 3/50
8/8 [=====] - 0s 9ms/step - loss: 0.8227 - sparse_categorical_accuracy: 0.7158 - val_loss: 0.7405 - val_sparse_categorical_accuracy: 0
Epoch 4/50
8/8 [=====] - 0s 8ms/step - loss: 0.7441 - sparse_categorical_accuracy: 0.7479 - val_loss: 0.6878 - val_sparse_categorical_accuracy: 0
Epoch 5/50
8/8 [=====] - 0s 7ms/step - loss: 0.7027 - sparse_categorical_accuracy: 0.7853 - val_loss: 0.6656 - val_sparse_categorical_accuracy: 0
Epoch 6/50
8/8 [=====] - 0s 9ms/step - loss: 0.6705 - sparse_categorical_accuracy: 0.8045 - val_loss: 0.6415 - val_sparse_categorical_accuracy: 0
Epoch 7/50
8/8 [=====] - 0s 7ms/step - loss: 0.6652 - sparse_categorical_accuracy: 0.8205 - val_loss: 0.6220 - val_sparse_categorical_accuracy: 0
Epoch 8/50
8/8 [=====] - 0s 7ms/step - loss: 0.6444 - sparse_categorical_accuracy: 0.8226 - val_loss: 0.5937 - val_sparse_categorical_accuracy: 0
Epoch 9/50
8/8 [=====] - 0s 7ms/step - loss: 0.6206 - sparse_categorical_accuracy: 0.8568 - val_loss: 0.5899 - val_sparse_categorical_accuracy: 0
Epoch 10/50
8/8 [=====] - 0s 6ms/step - loss: 0.6219 - sparse_categorical_accuracy: 0.8355 - val_loss: 0.6007 - val_sparse_categorical_accuracy: 0
Epoch 11/50
8/8 [=====] - 0s 7ms/step - loss: 0.6353 - sparse_categorical_accuracy: 0.8226 - val_loss: 0.5927 - val_sparse_categorical_accuracy: 0
```

Figure 8: Training the NN

	precision	recall	f1-score	support
0	0.91	0.88	0.90	113
1	0.81	0.80	0.81	111
2	0.69	0.93	0.79	76
3	0.99	0.76	0.86	102
accuracy			0.84	402
macro avg	0.85	0.85	0.84	402
weighted avg	0.86	0.84	0.84	402

Figure 9: Classification Report for NN

4. Decision Tree Classifier

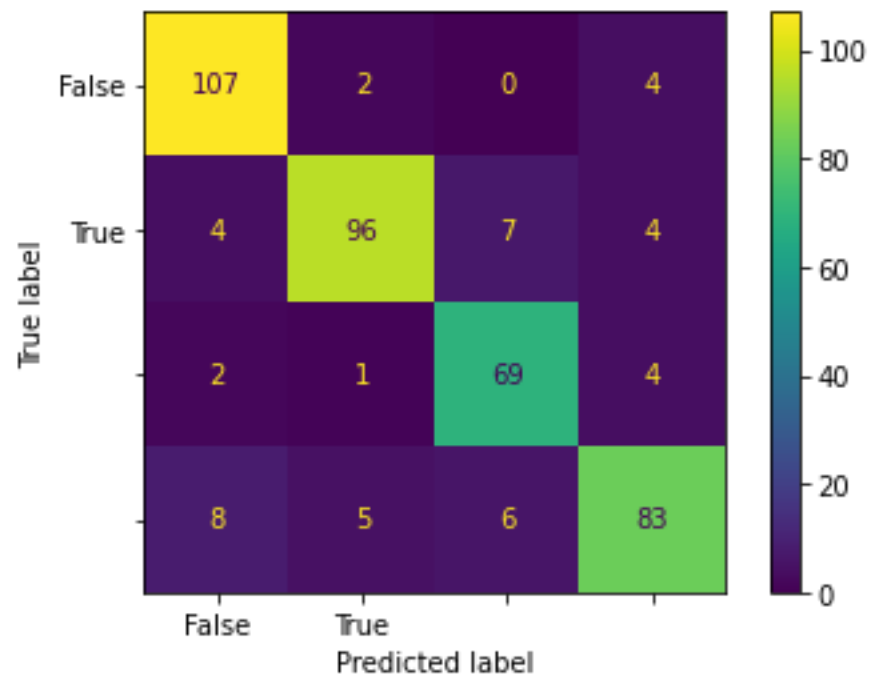


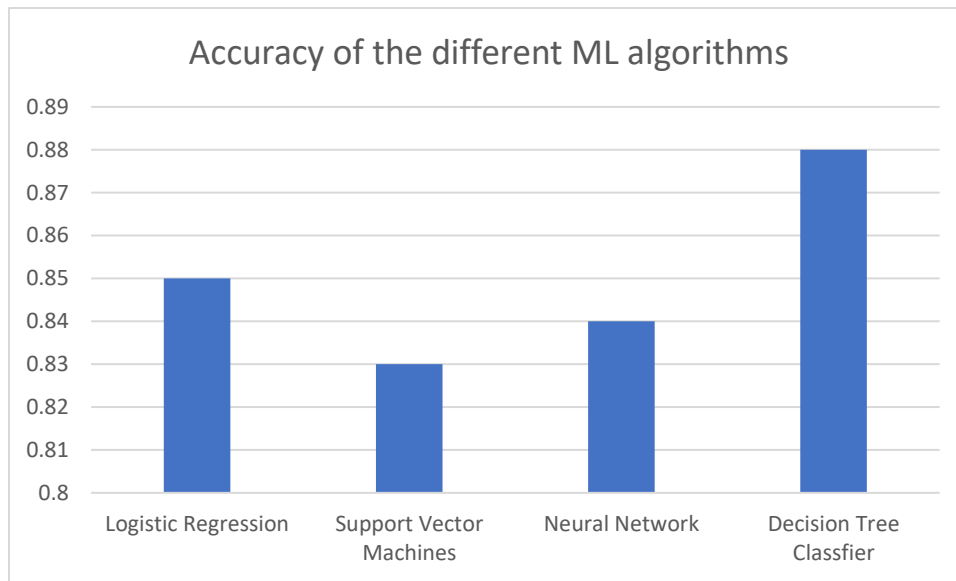
Figure 10:Confusion Matrix for Decision Tree Classifier

	precision	recall	f1-score	support
0	0.88	0.95	0.91	113
1	0.92	0.86	0.89	111
2	0.84	0.91	0.87	76
3	0.87	0.81	0.84	102
accuracy			0.88	402
macro avg	0.88	0.88	0.88	402
weighted avg	0.88	0.88	0.88	402

Figure 11:Classification Report for Decision Tree Classifier

DISCUSSION AND CONCLUSION

Same dataset has trained by using four different Machine Learning (ML) models to identify the best model. Figure 12 depicts the ML algorithm and the accuracy with f1-score for each algorithm.



According to the results, Decision Tree Classifier(DTC) shows much accuracy of predicting the best plan for each customer. Among these 4 classifiers SVM shows the lowest accuracy of predicting the insurance plan.

As a conclusion, Ravindu can use DTC to predict the most suitable insurance plan for their customers and promote those insurance plan.

Future Works

Ravindu can create an API from the model and create a simple user interface when showing these results to the management as the progress of the work. Finally, this can be handover to the front office as well as a built in function in the system.