CS 513 KDD
PRUDENTIAL LIFE INSURANCE ASSESSMENT

GUIDED BY: PROF. KHASHAYAR DEHNAD

Team Members:

Shreya Mahadik (10467745)
Parth Ambalkar (10467986)
Pinak Pathak (10472891)
Aditi Duggal (10460663)

AGENDA

PROBLEM STATEMENT

DATASET

DATA PRE-PROCESSING (EDA)

MACHINE LEARNING MODELS

MODEL EVALUATION

CONCLUSION

PROBLEM STATEMENT

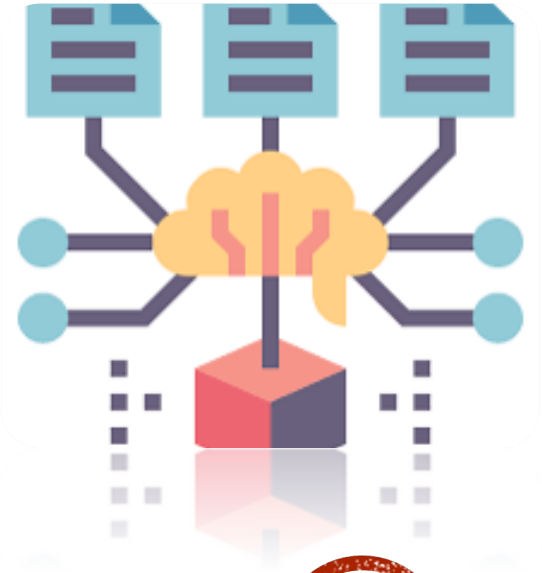
The process of life insurance purchase is complicated and involves risk analysis at multiple levels and various medical exams.

Prudential wants to make the process quicker, less labor intensive and automated.

By using predictive modeling, we are classifying the risk profiles of the customer based on their information.

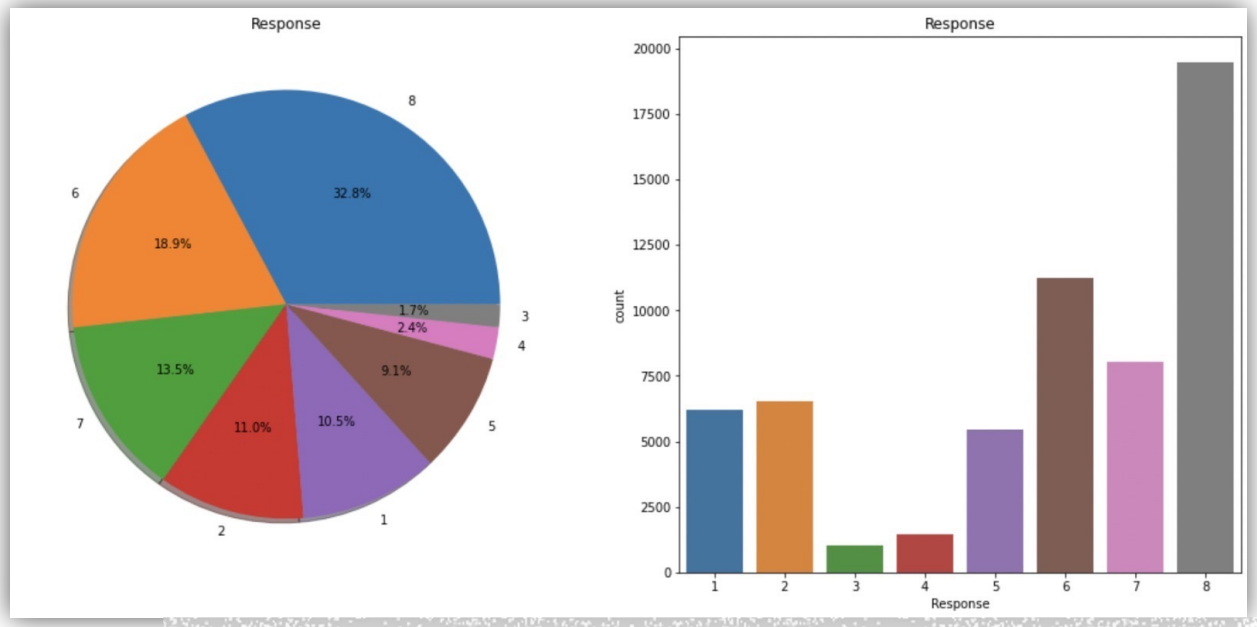
DATASET

Train.csv has 59k rows and 128 columns



DATA PREPROCESSING

EDA



Basic Information About The Dataset

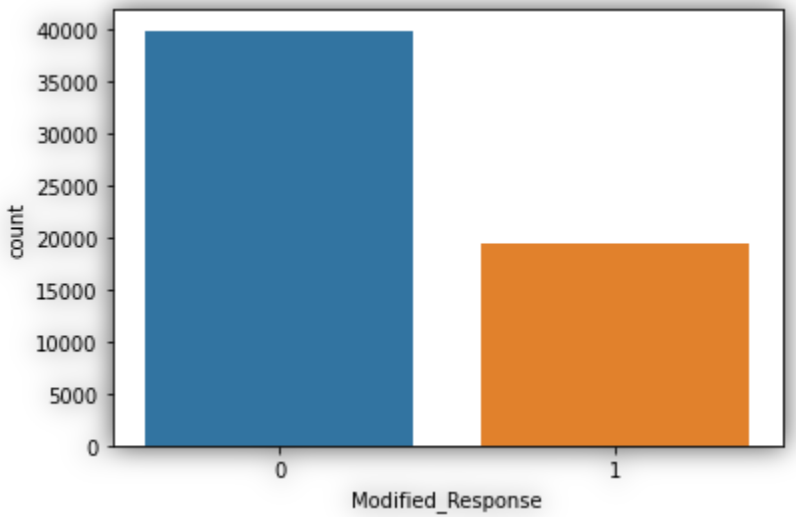
- In all, there are 8 target variables.
- Wherein the 8th variable has the highest COUNT.
- As this is a classification model the target variables were bifurcated into 2 classes:
  - **0 (1st -7th Feature)**
  - **1 (8th Feature)**

EXPLORATORY DATA ANALYSIS

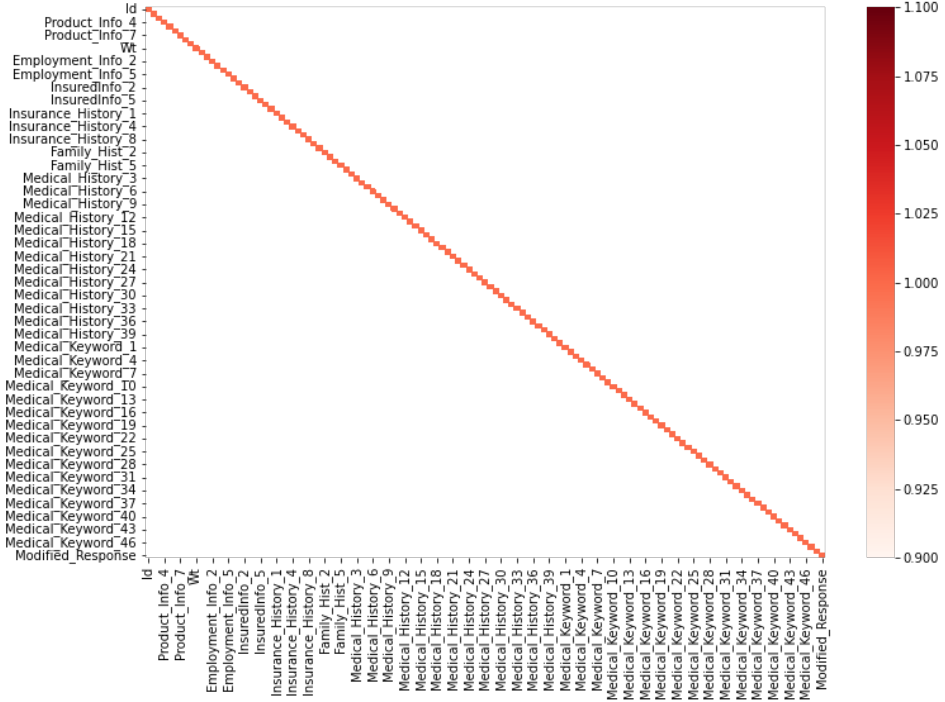
EXPLORATORY DATA ANALYSIS

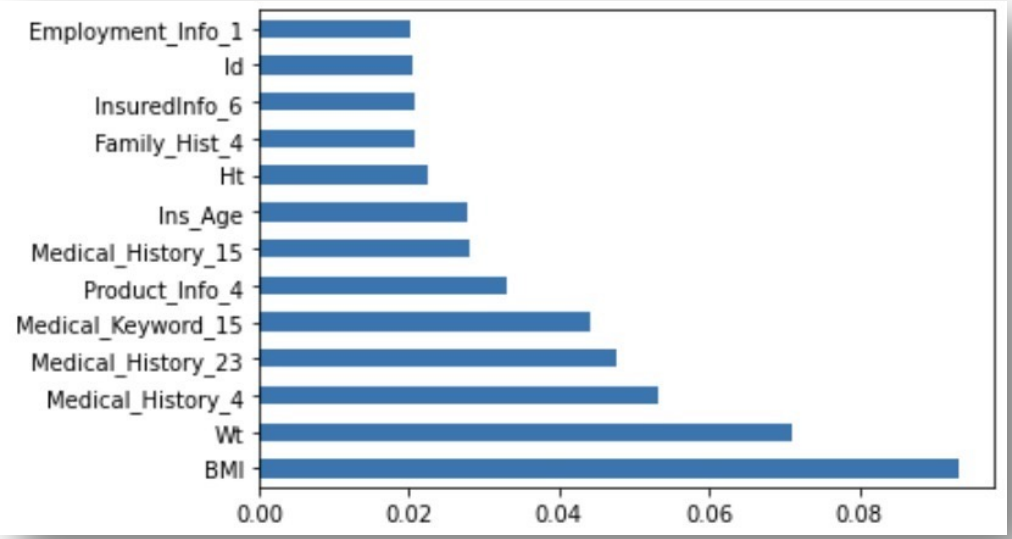
Here, data is still imbalanced, and the classification model classes are:

- **0(8th Feature)**
- **1(1st -7th Feature)**



CORRELATION MATRIX OF ENTIRE DATASET

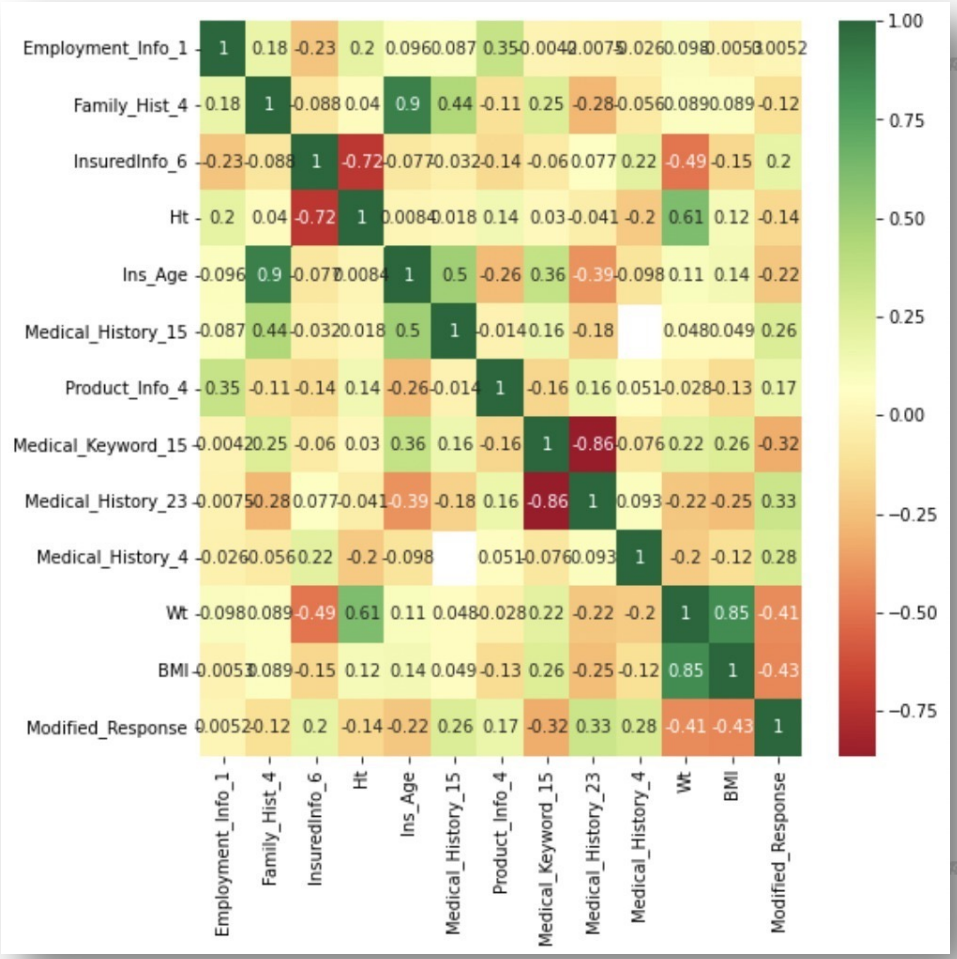




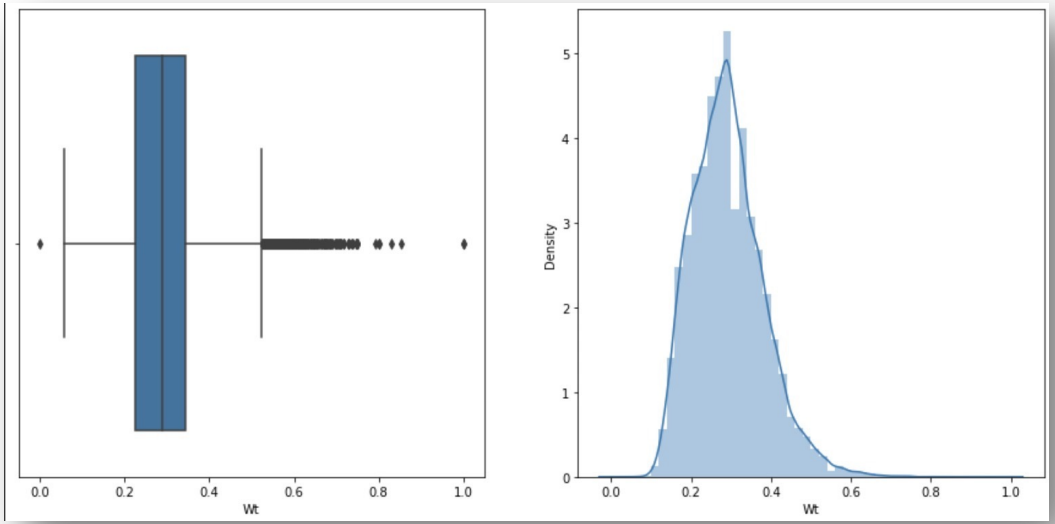
- Categorical and Numerical features were analyzed

- Out of the entire feature set, 12 best features were selected after the Correlation Matrix was found.

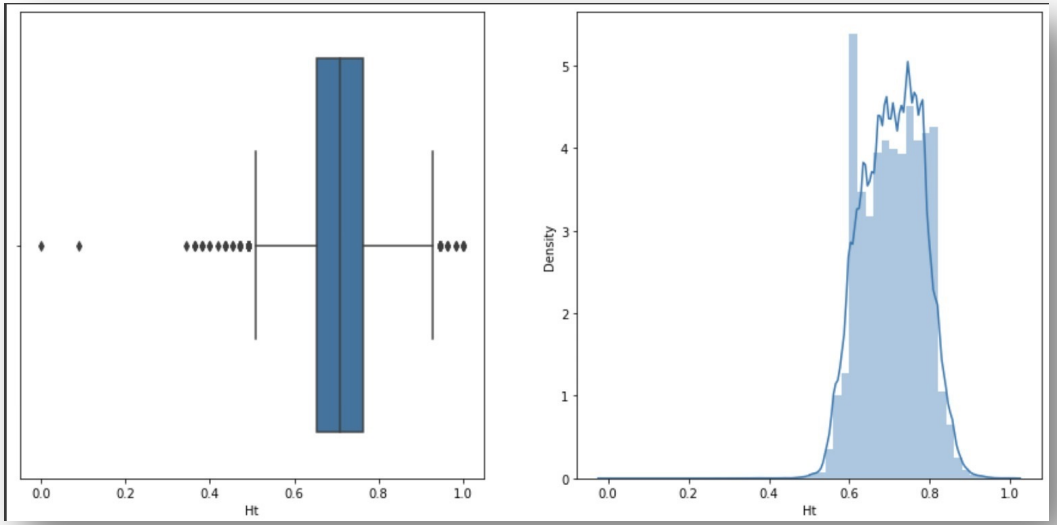
EXPLORATORY DATA ANALYSIS



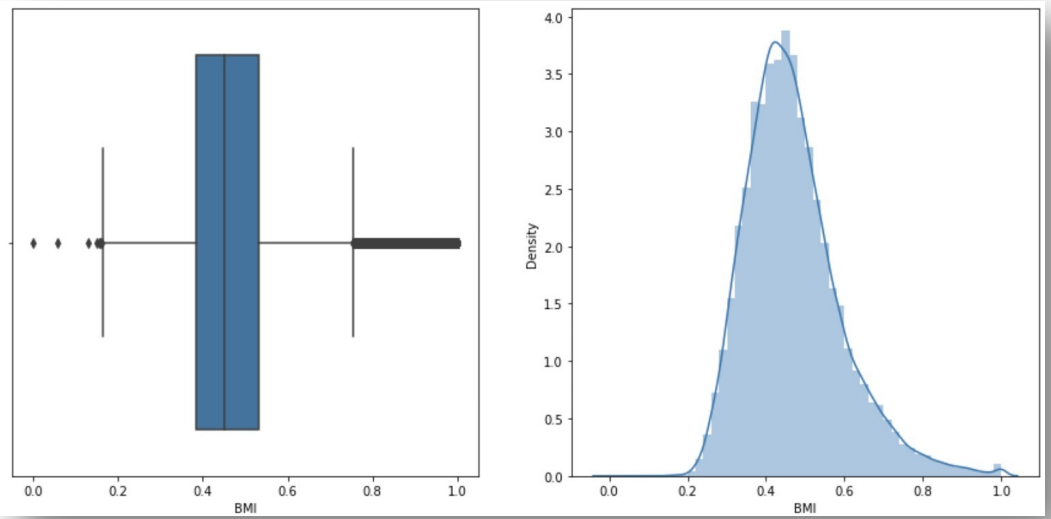
CORRELATION MATRIX OF 12 FEATURES



WEIGHT



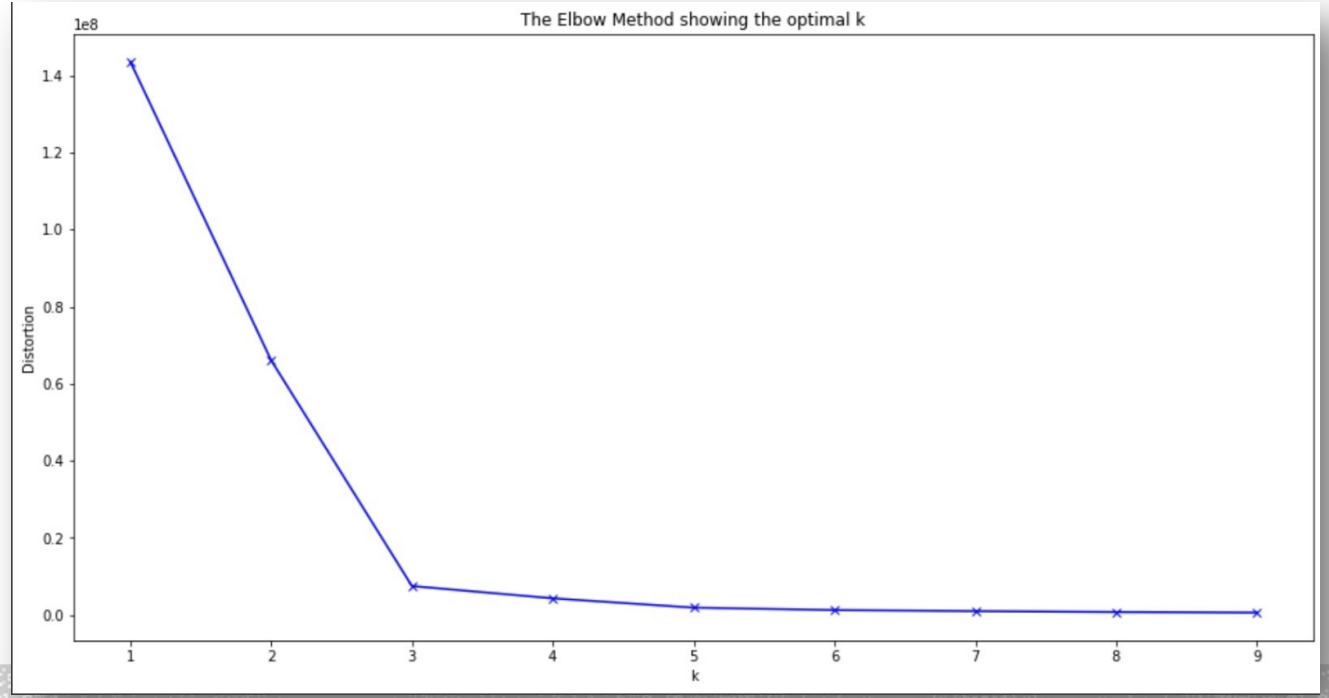
HEIGHT


0.0
0.2
0.4
0.6
0.8
1.0
BMI
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Density
0.0
0.2
0.4
0.6
0.8
1.0
BMI

BMI

CLASSIFICATION ALGORITHMS IMPLEMENTED

- KNN
- NAÏVE BAYES
- DECISION TREES-C5.0
- DECISION TREES-CART
- RANDOM FOREST
- ANN



ELBOW DIAGRAM: TO FIND THE IDEAL NUMBER OF CLUSTERS

ACCURACY

```
 The confusion matrix and accuracy score for n_neighbors =

[[6119 1900]
 [ 906 2952]]

 The accuracy score for n_neighbors =

0.7637450534646796
```

K NEAREST NEIGHBORS ALGORITHM

BEST PARAMETERS

```
BEST PARAMS: {'n_neighbors': 3}

0.836 (+/-0.03) for {'n_neighbors': 3}
0.829 (+/-0.041) for {'n_neighbors': 4}
0.829 (+/-0.028) for {'n_neighbors': 5}
```

ACCURACY

```
 The confusion matrix for Naive Bayes is =

[[5304 2715]
 [ 406 3452]]

 The accuracy score for Naive Bayes is =

0.7372232045129241
```

NAIVE BAYES ALGORITHM

ACCURACY

CART

```
 The confusion matrix for CART is =

[[6474 1545]
 [1290 2568]]

 The accuracy score for CART is =

0.7613033594342006
```

C 5.0

```
[[6467 1552]
 [1317 2541]]

 The confusion matrix for C5.0 is =


 The accuracy score for C5.0 is =

0.7584406836743285
```

DECISION TREE CLASSIFIER

ACCURACY

```
 The confusion matrix for no. of trees =

[[6510 1509]
 [ 759 3099]]

 The accuracy score for no. of trees =

0.8090426875473604
```

RANDOM FOREST CLASSIFIER

BEST PARAMETERS

```
BEST PARAMS: {'criterion': 'entropy', 'n_estimators': 200}

0.856 (+/-0.052) for {'criterion': 'entropy', 'n_estimators': 50}
0.857 (+/-0.047) for {'criterion': 'entropy', 'n_estimators': 100}
0.857 (+/-0.048) for {'criterion': 'entropy', 'n_estimators': 150}
0.858 (+/-0.047) for {'criterion': 'entropy', 'n_estimators': 200}
0.857 (+/-0.047) for {'criterion': 'entropy', 'n_estimators': 250}
0.856 (+/-0.05) for {'criterion': 'gini', 'n_estimators': 50}
0.857 (+/-0.049) for {'criterion': 'gini', 'n_estimators': 100}
0.858 (+/-0.048) for {'criterion': 'gini', 'n_estimators': 150}
0.858 (+/-0.048) for {'criterion': 'gini', 'n_estimators': 200}
0.858 (+/-0.047) for {'criterion': 'gini', 'n_estimators': 250}
```

ACCURACY

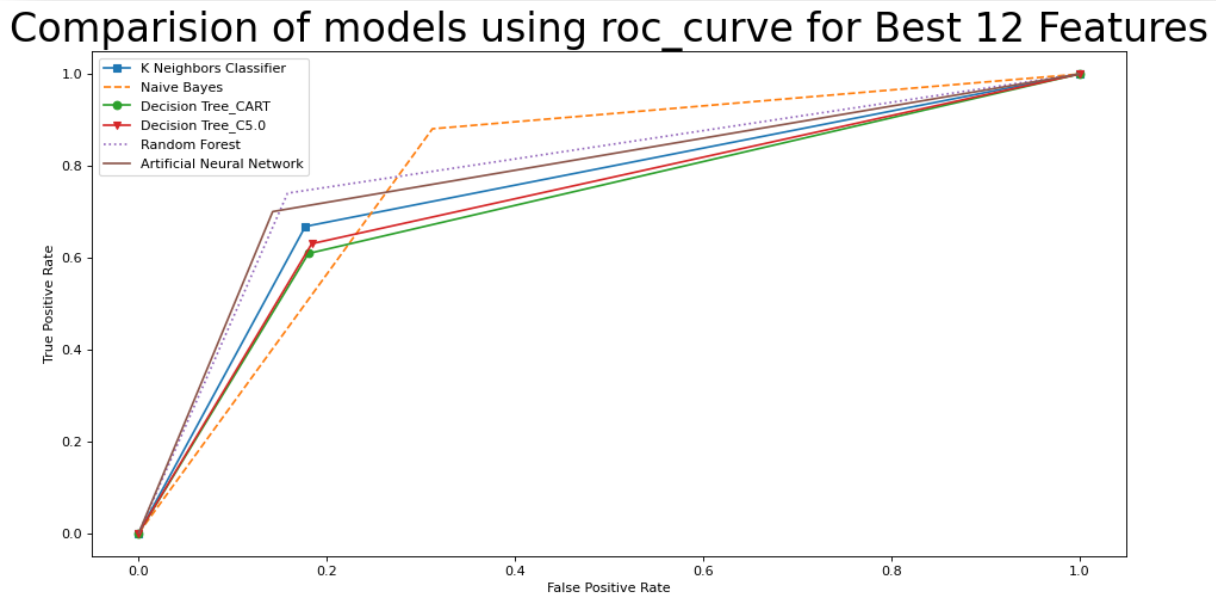
```
 The confusion Matrix for ANN

[[6072 1947]
 [ 467 3391]]

 The accuracy score for ANN

0.7967500210490864
```

ARTIFICIAL NEURAL NETWORK



COMPARISON OF MODELS USING ROC CURVE

SR.NO.	MODELS	MODEL ACCURACY	AUC
0	KNN	0.763745	0.764113
1	NAÏVE BAYES	0.737223	0.778097
2	CART	0.761303	0.736481
3	C5.0	0.758441	0.732546
4	RANDOM FOREST	0.809043	0.807544
5	ANN	0.796750	0.818077

COMPARISON OF MODELS BY THEIR ACCURACY AND AUC

WHICH MODEL TO CHOOSE?

CONCLUSION

Based on *Accuracy* as Evaluation Metric, ***RANDOM FOREST p***erforms well.

Based on *ROC/AUC* as Evaluation Metrics ***ANN*** performs best.

THANK YOU!