MACHINE LEARNING MODEL FOR PREDICTING TRAVEL INSURANCE

MUHAMMAD AULIA RENDY

USING PYTHON

ABOUT THE PROJECT

1. This project was carried out with the aim to build and find Machine learning model that have good result to predict whether someone will buy Travel Insurance packages or not.

2. Because in imbalanced data condition, F1 score is the proper metric to measure the model performance.

ANALYSIS FLOW

1. Data Undestanding
2. Data Cleaning
3. Exploratory Data Analysis
4. Modelling
    a. CatBoost
    b. Naïve Bayes
    c. Random Forest
5. Evaluation and Recommendation

DATA UNDERSTANDING

DATA UNDERSTANDING

The data is taken from Kaggle : Travel Insurance Prediction Data

It consist 1987 rows and 9 columns

The goal of this dataset is to predict whether a customer will be interested in buying travel insurance or not

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1987 entries, 0 to 1986
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Age                  1987 non-null   int64
 1   Employment Type      1987 non-null   object
 2   GraduateOrNot        1987 non-null   object
 3   AnnualIncome         1987 non-null   int64
 4   FamilyMembers        1987 non-null   int64
 5   ChronicDiseases      1987 non-null   int64
 6   FrequentFlyer        1987 non-null   object
 7   EverTravelledAbroad  1987 non-null   object
 8   TravelInsurance      1987 non-null   int64
dtypes: int64(5), object(4)
memory usage: 139.8+ KB
```

DATA UNDERSTANDING

Data dictionary

column	information
Age	Age Of The Customer
Employment Type	The Sector In Which Customer Is Employed
GraduateOrNot	Whether The Customer Is College Graduate Or Not
AnnualIncome	The Yearly Income Of The Customer In Indian Rupees[Rounded To Nearest 50 Thousand Rupees]
FamilyMembers	Number Of Members In Customer's Family
ChronicDisease	Whether The Customer Suffers From Any Major Disease Or Conditions Like Diabetes/High BP or Asthama,etc.
FrequentFlyer	Derived Data Based On Customer's History Of Booking Air Tickets On Atleast 4 Different Instances In The Last 2 Years[2017-2019].
EverTravelledAbroad	Has The Customer Ever Travelled To A Foreign Country[Not Necessarily Using The Company's Services]
TravelInsurance	Did The Customer Buy Travel Insurance Package During Introductory Offering Held In The Year 2019.

DATA CLEANING

DATA CLEANING

SPLIT DATASET

```
No. of training examples: 1390
No. of testing examples: 597
```

Before we do data cleaning, we have to split between data for training dan data for testing.

Allocation between train data and test data is 70:30

DUPLICATE CHECK

DATA TRAIN

```
Data dimension before duplicate handling:1390
Data dimension after duplicate handling:955
```

DATA TEST

```
Data dimension before duplicate handling:597
Data dimension after duplicate handling:500
```

DATA CLEANING

MISSING VALUE CHECK

DATA TRAIN

```
Age                    0
Employment Type        0
GraduateOrNot          0
AnnualIncome           0
FamilyMembers          0
ChronicDiseases        0
FrequentFlyer          0
EverTravelledAbroad    0
TravelInsurance        0
dtype: int64
```

MISSING VALUE CHECK

DATA TEST

```
Age                    0
Employment Type        0
GraduateOrNot          0
AnnualIncome           0
FamilyMembers          0
ChronicDiseases        0
FrequentFlyer          0
EverTravelledAbroad    0
TravelInsurance        0
dtype: int64
```

DATA CLEANING FEATURE ENCODING

COLUMN SELECTED : [EMPLOYMENT TYPE, GRADUATEORNOT, FREQUENTFLYER, EVERTRAVELLEDABROAD]

BEFORE

Age	Employment Type	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer	EverTravelledAbroad	TravelInsurance
34	Government Sector	Yes	1300000	7	0	No	No	1
31	Private Sector/Self Employed	Yes	1250000	7	0	No	No	0
29	Private Sector/Self Employed	Yes	1200000	5	1	No	No	0
28	Private Sector/Self Employed	Yes	600000	3	0	No	No	0
26	Private Sector/Self Employed	Yes	500000	8	0	No	No	0

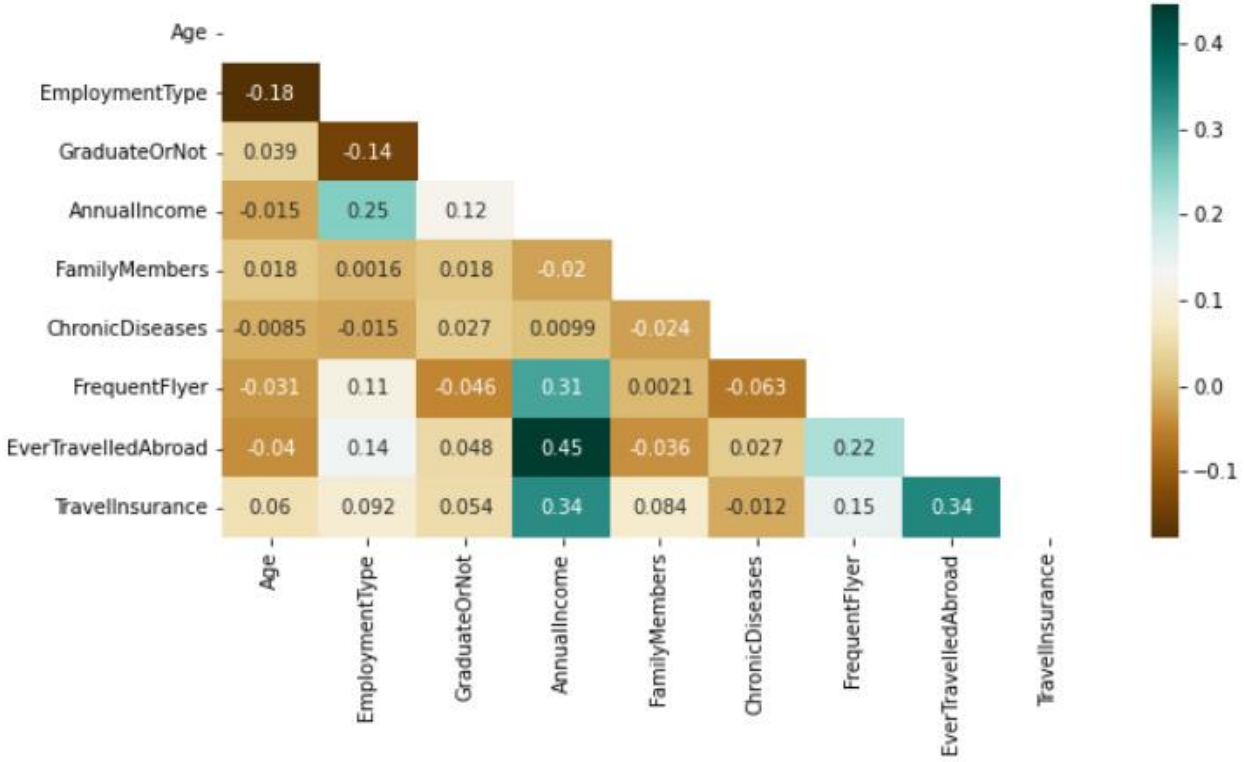
AFTER

Age	EmploymentType	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer	EverTravelledAbroad	TravelInsurance
34	0	1	1300000	7	0	0	0	1
31	1	1	1250000	7	0	0	0	0
29	1	1	1200000	5	1	0	0	0
28	1	1	600000	3	0	0	0	0
26	1	1	500000	8	0	0	0	0

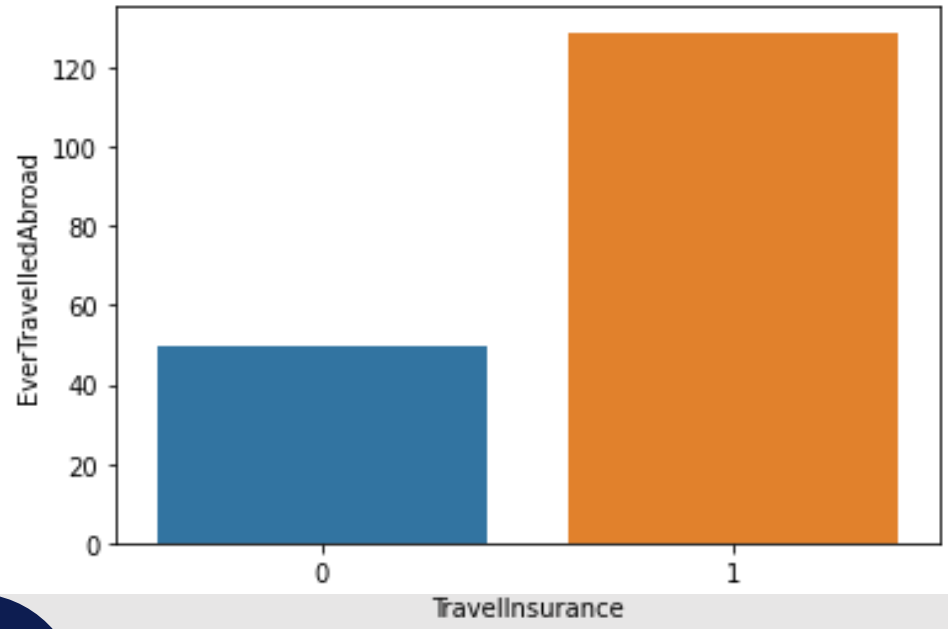
EXPLORATORY DATA ANALYSIS

EXPLORATORY DATA ANALYSIS

CORRELATION BETWEEN COLUMNS



Most of people that ever travelled aboard tend to buy for travel insurance packages

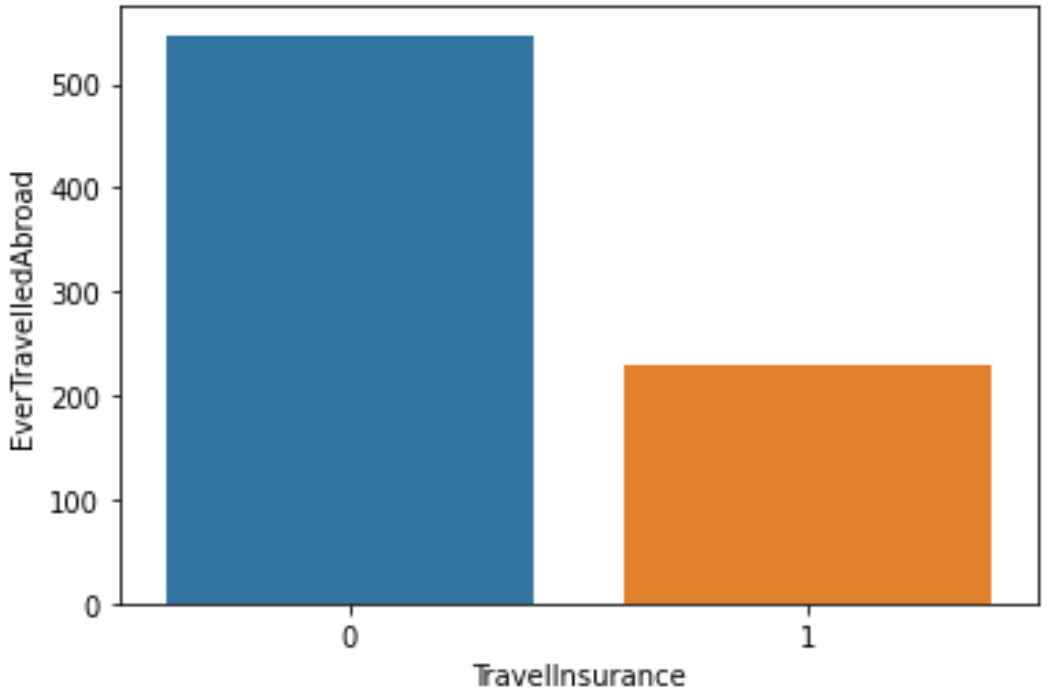


EVER TRAVELLED ABROAD VS TRAVEL INSURANCE

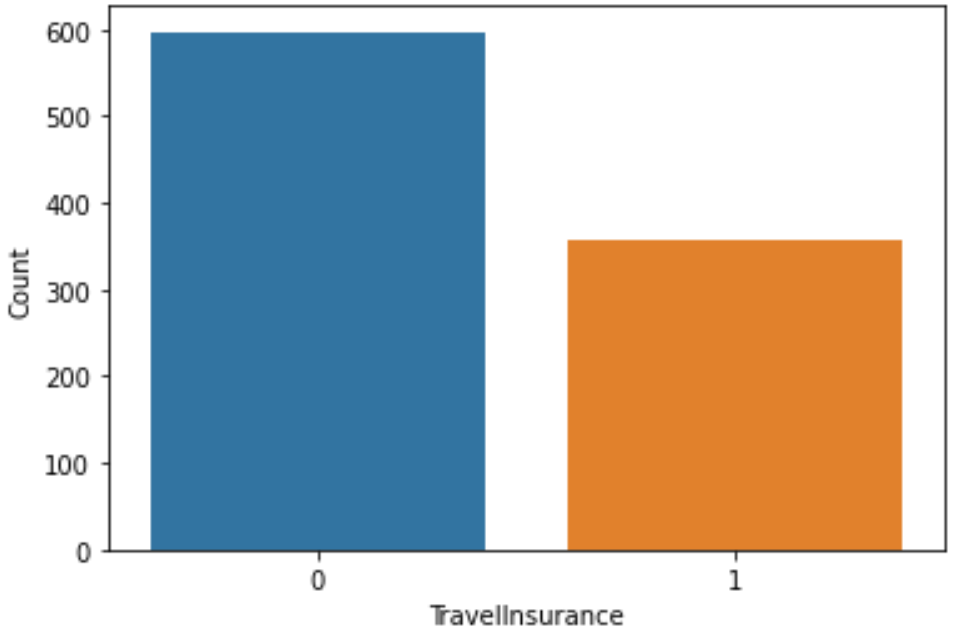
All the predictors are not highly correlated with the data target(y)

EXPLORATORY DATA ANALYSIS

EVER TRAVELLED ABROAD VS TRAVEL INSURANCE

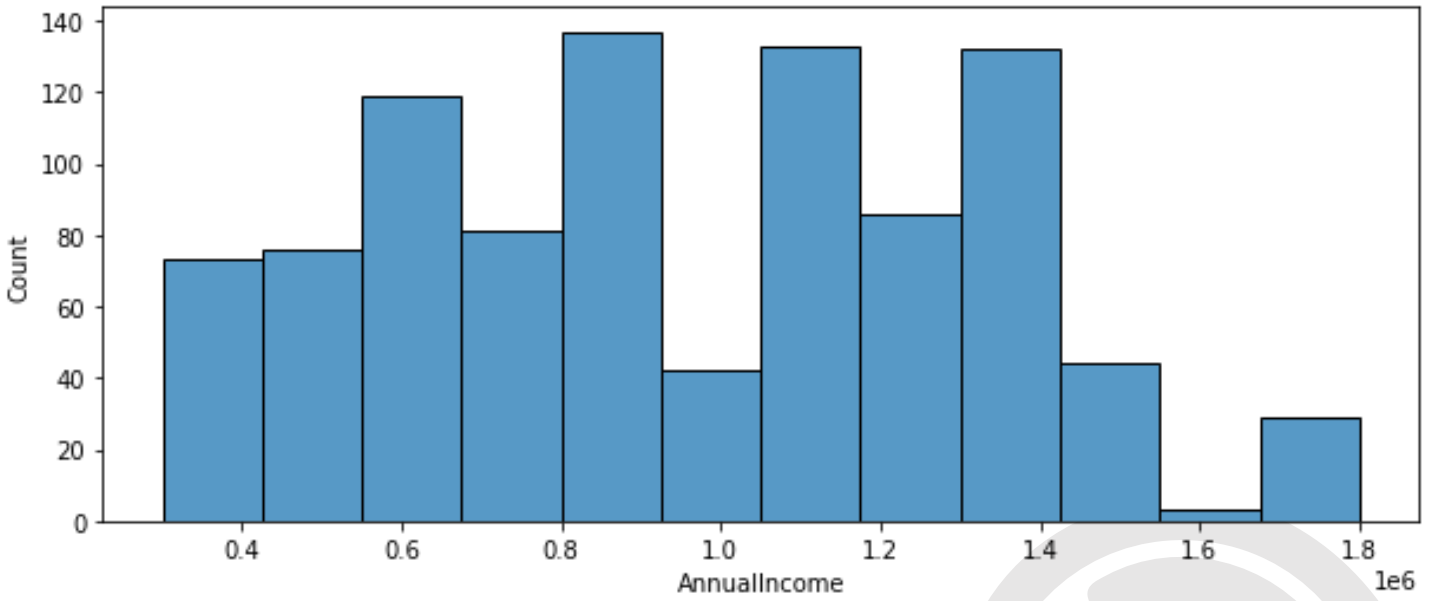


Most of people that never travelled aboard,only a few buy travel insurance packages



TRAVEL INSURANCE DISTRIBUTION DATA

EXPLORATORY DATA ANALYSIS

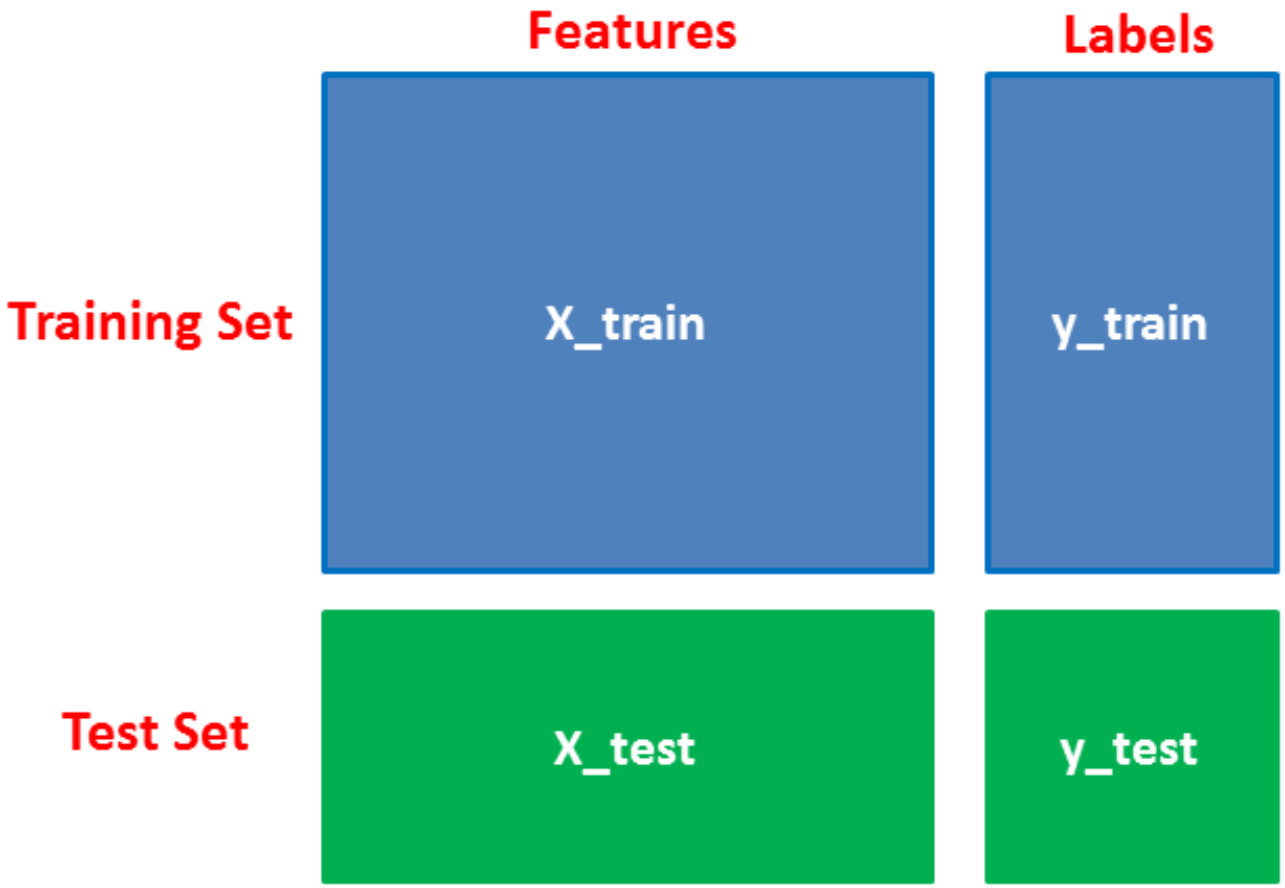


Annual Income (in rupee) distribution

MODELLING


BOARDING PASS
BOARDING PASS
TRAVEL INSURANCE
Date of Birth:
Male [ ]
Female [ ]

SPLITTING TRAIN AND TEST DATA



MACHINE LEARNING MODEL

CATBOOST

	precision	recall	f1-score	support
0	0.77	0.96	0.85	309
1	0.89	0.53	0.67	191
accuracy			0.80	500
macro avg	0.83	0.75	0.76	500
weighted avg	0.82	0.80	0.78	500

True Positives: 102
False Positives: 12
True Negatives: 297
False Negatives: 89

Because there is an imbalance of values on the label(y), it can be seen that the F1 score (0 & 1) has a very far difference

	precision	recall	f1-score	support
0	0.78	0.77	0.77	309
1	0.63	0.64	0.64	191
accuracy			0.72	500
macro avg	0.71	0.71	0.71	500
weighted avg	0.72	0.72	0.72	500

True Positives: 123
False Positives: 71
True Negatives: 238
False Negatives: 68

The results after resampling and setting the threshold, it can be seen that the F1 score (0 & 1) is better and the difference is not too far

CATBOOST WITH RESAMPLING AND SET THRESHOLD

Train

F1 score: 0.8275027097940557
Accuracy score: 0.8261780104712042

Test

F1 score: 0.7224050681026292
Accuracy score: 0.722

MACHINE LEARNING MODEL

NAÏVE BAYES

	precision	recall	f1-score	support
0	0.72	0.95	0.82	309
1	0.83	0.40	0.54	191
accuracy			0.74	500
macro avg	0.77	0.67	0.68	500
weighted avg	0.76	0.74	0.71	500

True Positives: 76
False Positives: 16
True Negatives: 293
False Negatives: 115

Because there is an imbalance of values on the label(y), it can be seen that the F1 score (0 & 1) has a very far difference

	precision	recall	f1-score	support
0	0.76	0.68	0.72	309
1	0.56	0.66	0.61	191
accuracy			0.67	500
macro avg	0.66	0.67	0.66	500
weighted avg	0.69	0.67	0.68	500

True Positives: 126
False Positives: 98
True Negatives: 211
False Negatives: 65

The results after resampling and setting the threshold, it can be seen that the F1 score (0 & 1) is better and the difference is not too far

NAÏVE BAYES WITH RESAMPLING AND SET THRESHOLD

Train

F1 score: 0.6671916466263458
Accuracy score: 0.6628272251308901

Test

F1 score: 0.6777665739882607
Accuracy score: 0.674

MACHINE LEARNING MODEL

RANDOM FOREST

	precision	recall	f1-score	support
0	0.76	0.85	0.80	309
1	0.70	0.57	0.63	191
accuracy			0.74	500
macro avg	0.73	0.71	0.72	500
weighted avg	0.74	0.74	0.74	500

True Positives: 109
False Positives: 47
True Negatives: 262
False Negatives: 82

Because there is an imbalance of values on the label(y), it can be seen that the F1 score (0 & 1) has a very far difference

	precision	recall	f1-score	support
0	0.77	0.76	0.77	309
1	0.62	0.64	0.63	191
accuracy			0.71	500
macro avg	0.70	0.70	0.70	500
weighted avg	0.72	0.71	0.71	500

True Positives: 122
False Positives: 74
True Negatives: 235
False Negatives: 69

The results after resampling and setting the threshold, it can be seen that the F1 score (0 & 1) is better and the difference is not too far

RANDOM FOREST WITH RESAMPLING AND SET THRESHOLD

Train

F1 score: 0.8806582620250663
Accuracy score: 0.8795811518324608

Test

F1 score: 0.7146811504398666
Accuracy score: 0.714

EVALUATION METRIC AND RECOMMENDATION

EVALUATION METRIC

Model Result metric

```
========CatBoost========
F1 score: 0.7224050681026292
Accuracy score: 0.722

========Naive Bayes========
F1 score: 0.6777665739882607
Accuracy score: 0.674

========Random Forest========
F1 score: 0.7146811504398666
Accuracy score: 0.714
```

After the imbalance problem was solve d by resampling and set threshold, it was found that the machine learning catboost model gave better F1 score results than the others 2 models.

Because in imbalanced data condition, F1 score is the metric to measure the model performance.

RECOMMENDATION

In order to produce good prediction results, here are recommendations for the travel insurance dataset:

1. It is necessary to add features that have a strong correlation with the target data to be able to improve prediction results using machine learning models
2. It is necessary to increase the number of data/observation, so the data will has a balance of values in the target column(y).

THANK YOU