Analysis and Predictive Modelling

For

Profusion Data Academy Technical Test

Assignment: Predict Credit card default and assess presence of algorithmic bias in Bank Dataset

Adhil Vengat
av21221@essex.ac.uk
Github link:
https://github.com/ADHIL-VENGAT/Profussion

March 11, 2022

Word Count: 1445

TABLE OF CONTENTS

Overview

This is predictive modelling project on bank marketing dataset , to carry out extensive exploratory data analysis and visualisation with the help of data analysis and machine learning techniques. The aim of the project is to predict Credit card default and assess presence of algorithmic bias in Bank Marketing.

Also the following prediction and analysis are proposed:

1. To predict and analyse the housing approval based on the age and profession
2. To predict and analyse the loan approval based on the age and profession

Investigation About Data

The dataset is provided from Kaggle(https://www.kaggle.com/shiv28/bank-marketing-slt/data?select=bank-additional-full.csv) which contains 41188 rows and 21 features. The dataset is targeted to determine the impact of call campaign in subscription of existing and new customers. It had no missing values at first glance but had many values labelled as "unknown", which were considered as missing values. There were many irrelevant features which had no affect or relation to the target "credit card default". Almost every features related to campaign call such as 'duration','pdays','previous','contact','month','day_of_week','pdays','poutcome' and 'y'. All these features were dropped due to their insignificance to target. The dataset was lacking features like 'salary', 'credit history' ,'savings' which could add on more detail to the model. The targets like loan , credit default, fraud are prone to imbalanced data set. From the primary analysis the dataset is found to be highly imbalanced.

Libraries

For this investigation following libraries are mainly used:

- Pandas
- Matplotlib

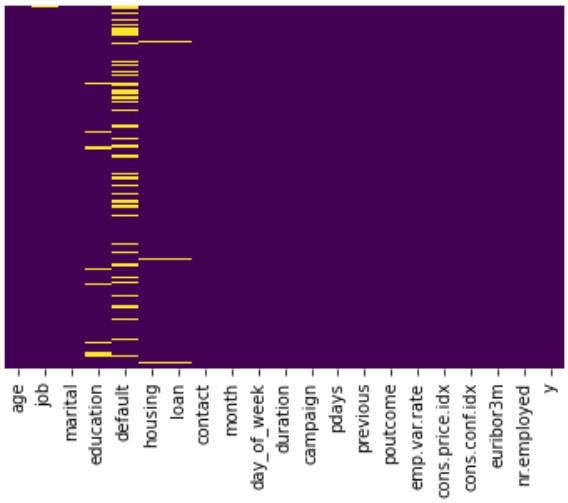
- Sklearn
- Seaborn
- dtale
- Pylab
- Scipy

Pre-processing

It is one of the most important part of analysis, dtale library was used to get an quick overview of the dataset

Exploratory data analysis is carried out on data to plot and visualize:

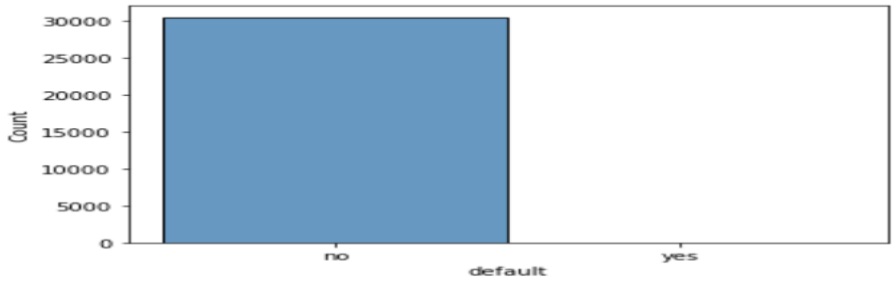
1) Missing Values : There were 10700 rows containing missing value labelled as ‘uknown’ which were dropped in which 80% of them were from target.



2) Checking data for Imbalance: The data is observed to be highly imbalanced containing almost 99.99% of same target value which is 'No credit default '.

```
no      30485
yes         3
Name: default, dtype: int64
```

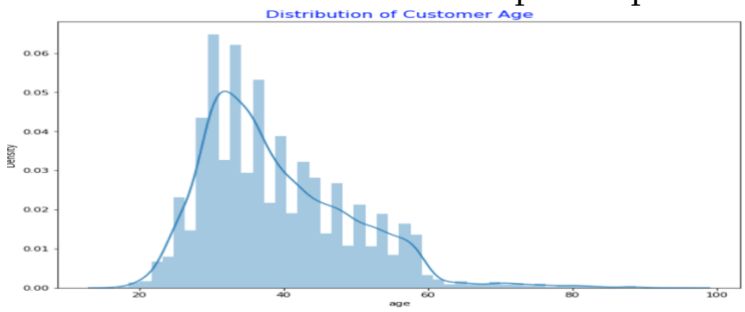
```
sns.histplot(df_unknown_dropped['default'])
<AxesSubplot:xlabel='default', ylabel='Count'>
```



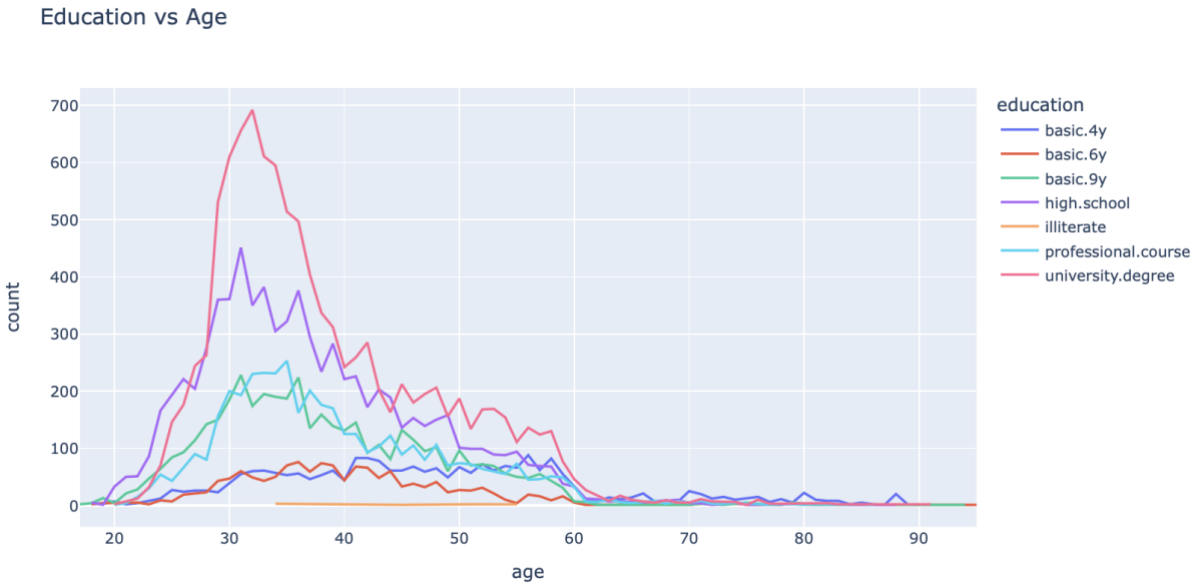
The dataset is highly imbalanced with a lot of 0 values

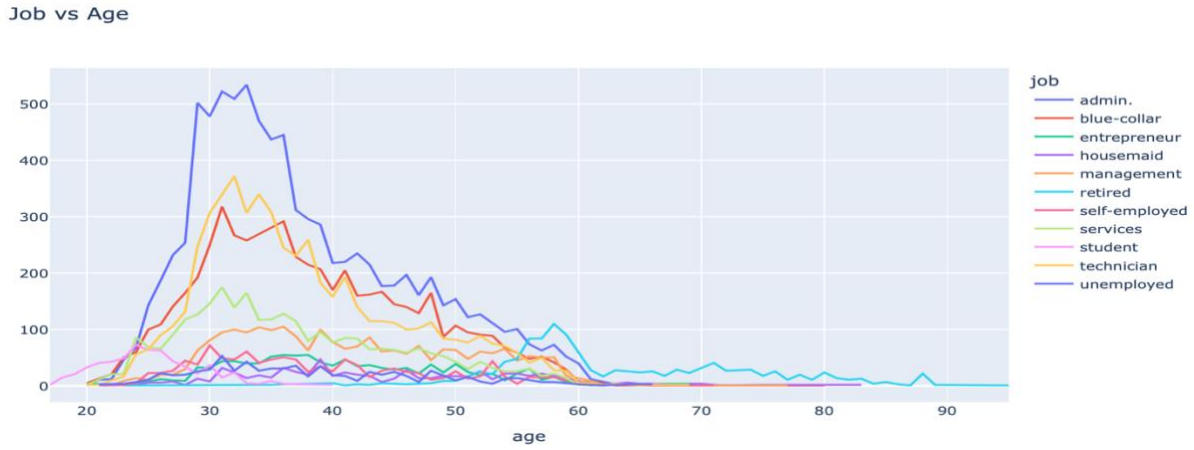
3) Analysing Numerical Variables and categorical variables:

- All of the Numerical Variables where analysed for their Discrete and continuous characteristics with help of distplot.

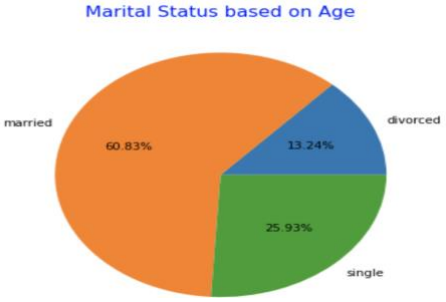


- The distribution of age vs job and age vs education is plotted for a better understanding of their density distribution

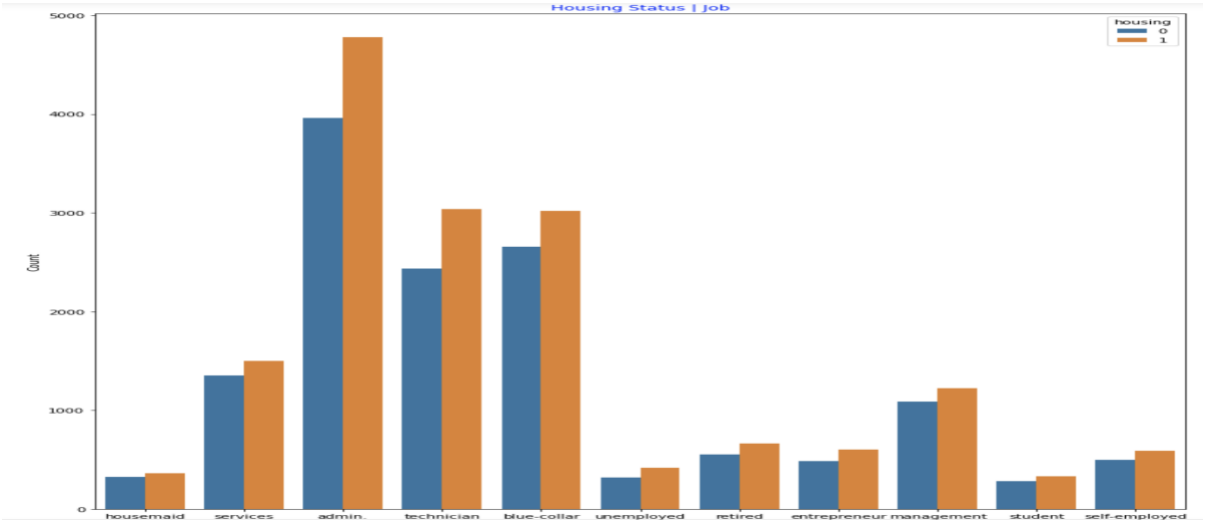


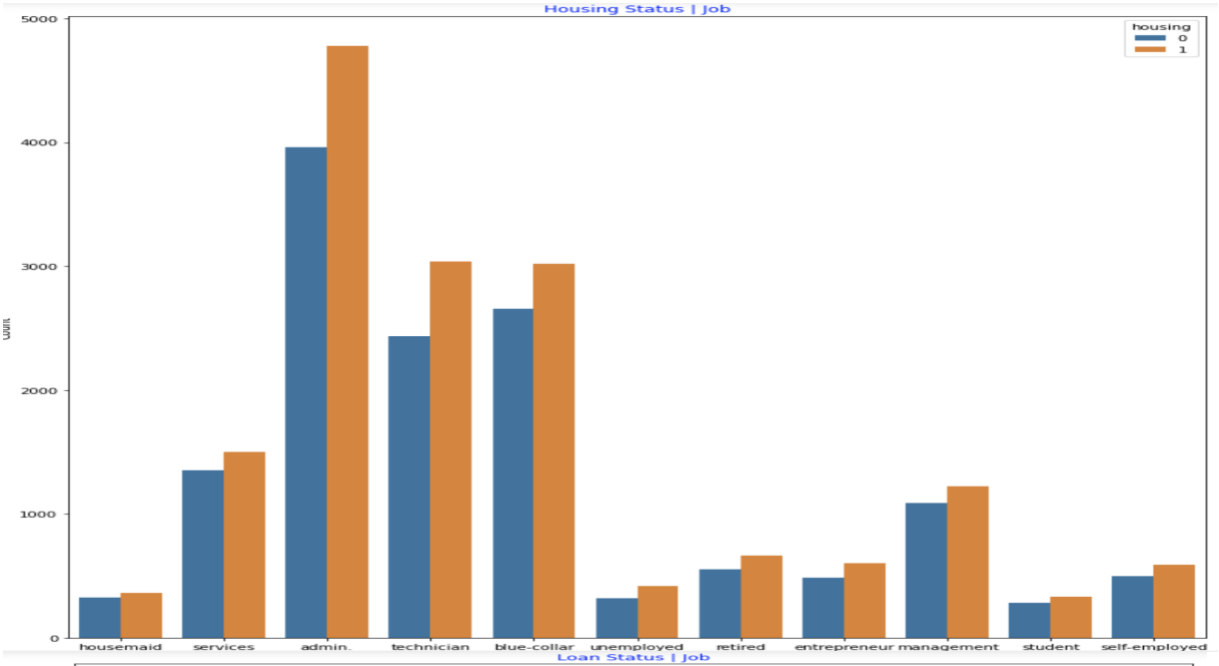


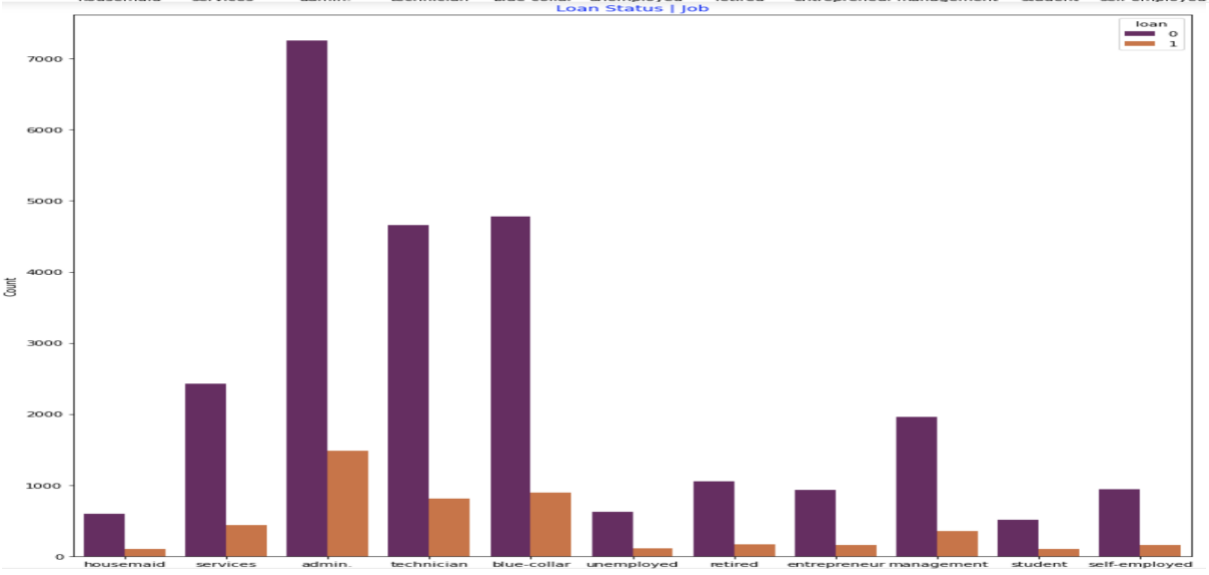
- Marital status is analysed with help of pie chart.

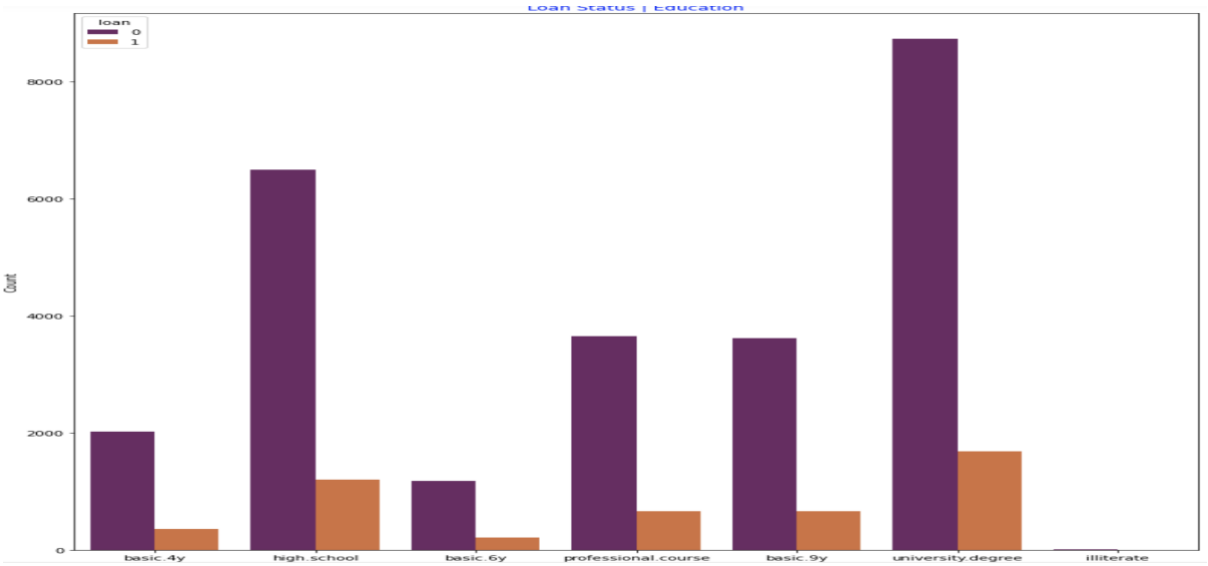


- The housing and loan feature against job and education is plotted to analyse the importance of those feature on bank decisions

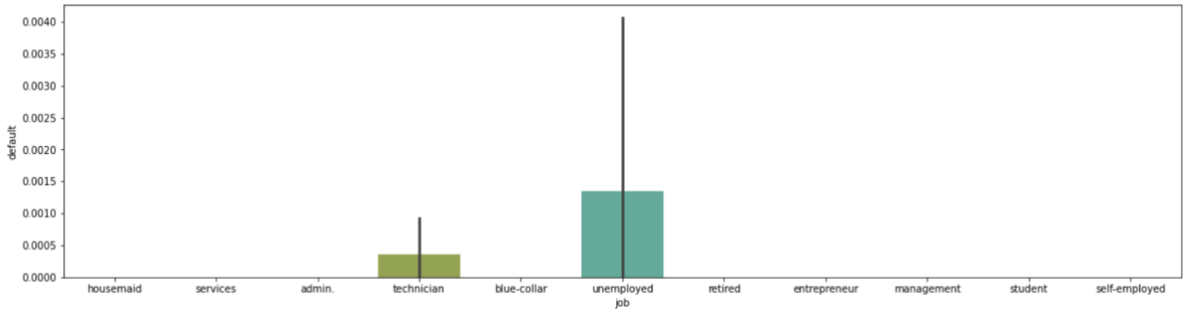

Housing Status | Job
housing
0
1
Count
5000
4000
3000
2000
1000
0
housemaid
services
admin.
technician
blue-collar
unemployed
retired
entrepreneur
management
student
self-employed


Housing Status | Job
housing
0
1
Count
5000
4000
3000
2000
1000
0
housemaid
services
admin.
technician
blue-collar
unemployed
retired
entrepreneur
management
student
self-employed

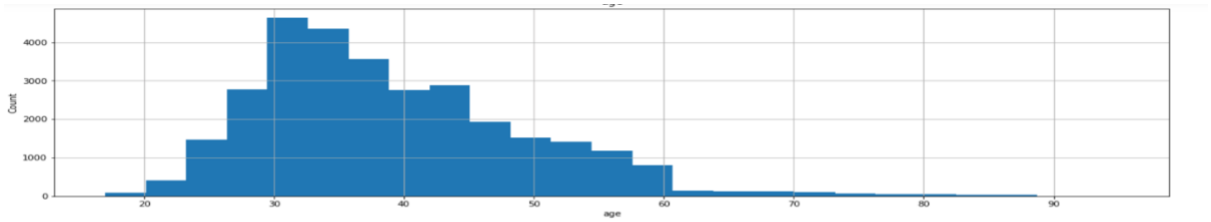

Loan Status | Job
loan
0
1
Count
7000
6000
5000
4000
3000
2000
1000
0
housemaid
services
admin.
technician
blue-collar
unemployed
retired
entrepreneur
management
student
self-employed



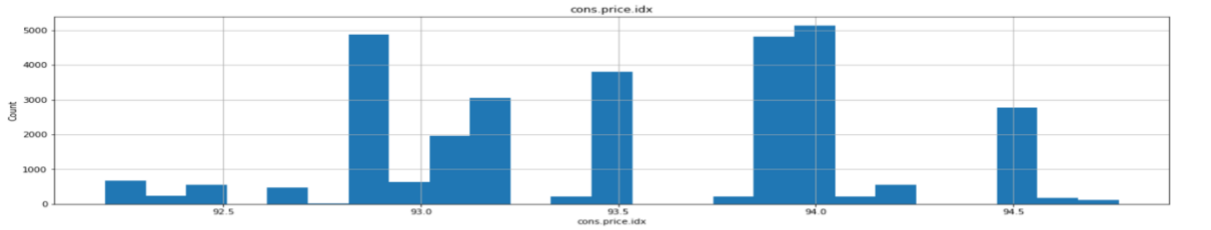
- It is observed that job can be ranked as most highly paid professional are likely to receive the benefits such as loan , housing and so on.
- In 'job' Feature ;including 'unemployed' and 'technician' are the only people who was categorised in credit card default.



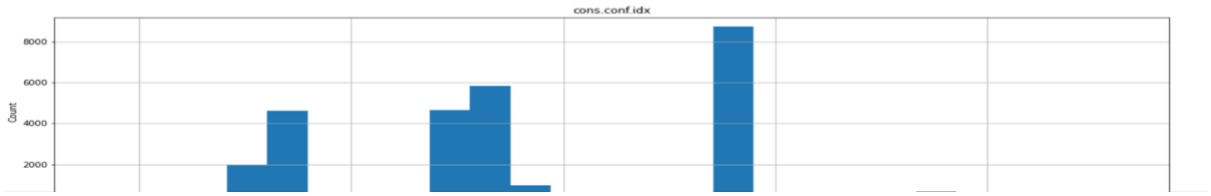
- Analysed if the skewed can be transformed to gaussian distribution using boxcox. It is observed that age can be transformed if needed. If skew score is close to 0 it indicated the closeness of the distribution towards normal distribution.

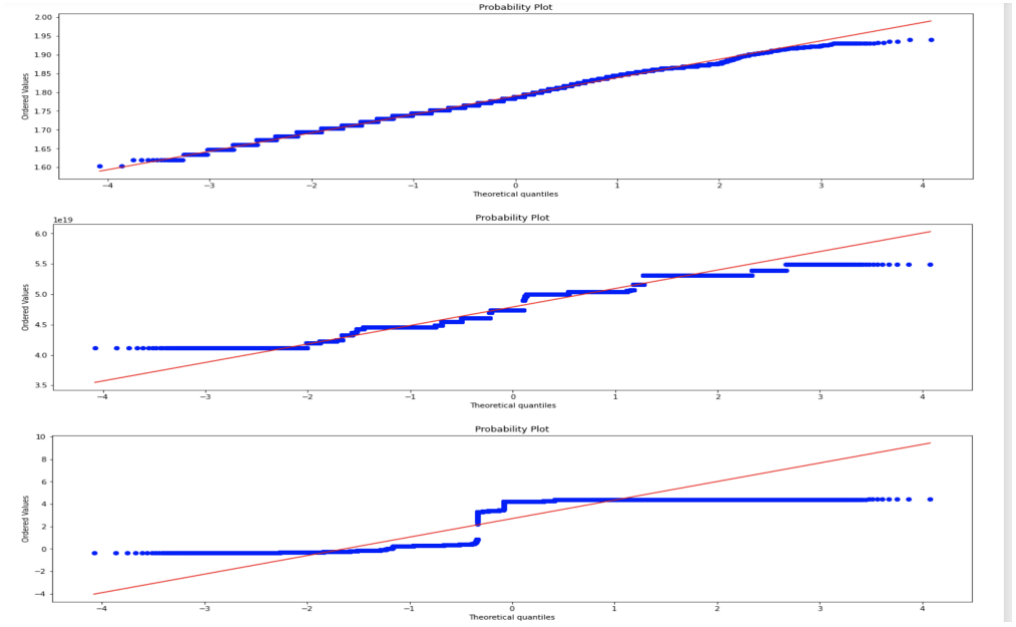


age skew score ----> 0.9801618327451445



cons.price.idx skew score ----> -0.11871890873194521

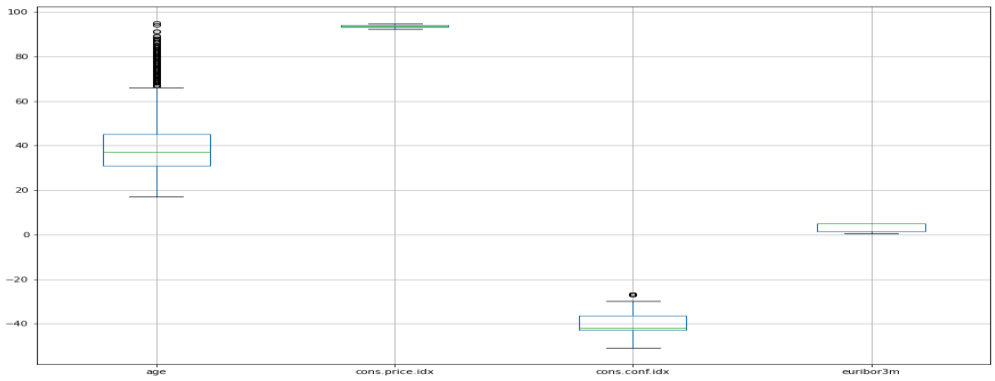


- 

4) Categorical Variables : Categorical features were analysed with help of histogram. And skew score to analyse its type of distribution so that in future we can try converting skewed to normal distribution with help normalisation techniques like boxcox.

5) Outliers:

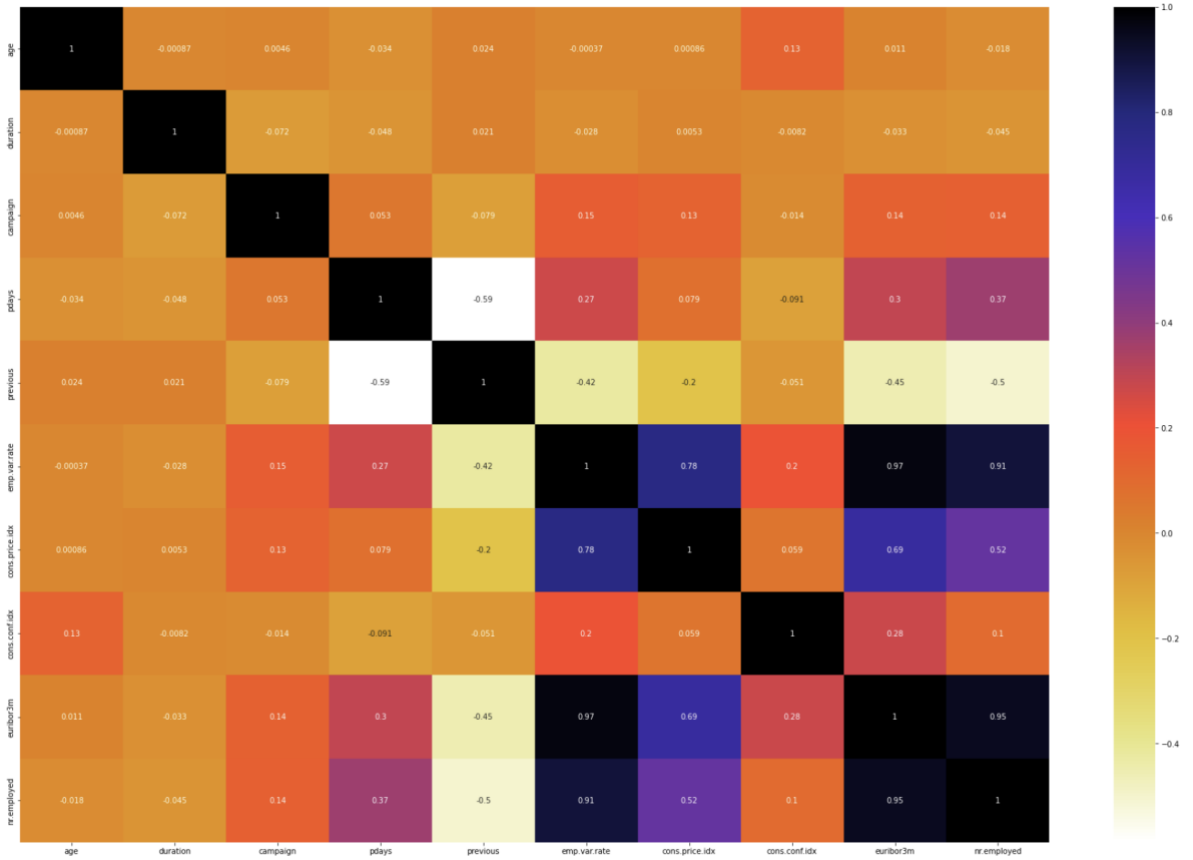
- Out of the continuous features age has few outliers .



Boxplot of outliers in continuous features

6) Correlation:

- Pearson correlation was plotted to analyse the correlation between continuous features. It was observed that ‘eubour3m’ and ‘nr.employed’ are highly correlated but nr.employed has discrete characteristics so they are not considered as highly correlated.



FEATURE ENGINEERING :

- Missing values('unknown') are dropped.
- **The target 'default' is categorical and is converted to binary using dummies.**
- There were many irrelevant features which had no affect or relation to the target "credit card default". Almost every features related to campaign call such as 'duration','pdays','previous','contact','month','day_of_week','pdays','poutcome' and 'y'. All these features were dropped on assuming their insignificance to target
- 'Job' was considered as ordinal categorical feature encoded and was ranked using rankmap.
- Categorical features such as are encoded using one hot encoder
- Positive skewed feature can be transformed into gaussian distribution for enhancing learning.
- Outliers of normal and skewed features having high skewness score are replaced with upper and lower bridge values found using inter quantile range.
- The discrete and continuous data is normalized using minmax scaler()

Handling the Imbalance in data frame

- The dataset is balanced using random oversampling technique.

  Random over sampling : The random oversampling method operates by replicating the randomly selected set of examples from the minority class so that the majority class does not have an overbearing presence during the training process
- The data was fairly balanced to almost 90%.

Train and Test Split

- The data was split into 70% of train data and 30% of test data. Which was stratified along the target. With a random state of 50.
- The train data is split into X set and y set ,where X set contains all the data after feature engineering except the Labelled Target feature and y set only containing the target data.
- The dataset X,y is randomly split into 70% of train data in X_train and y_train and the rest 30% to X_test and Y_test.

Modelling and Evaluation

1. **Random forest classifier**: It has reduced overfitting and it is more accurate than the decision tree in most of the cases .It eliminates the problems faced by single decision tree approach. Along with grid search cv the better results were found. The parameters were tuned to give the best fit. The count false positive dropped significantly after gridsearchcv from 660 to 310.

```
[[8836  310]
 [   0    1]]
0.9661091068109763
              precision    recall  f1-score   support

           0       1.00      0.97      0.98      9146
           1       0.00      1.00      0.01         1

    accuracy                           0.97      9147
   macro avg       0.50      0.98      0.49      9147
weighted avg       1.00      0.97      0.98      9147
```

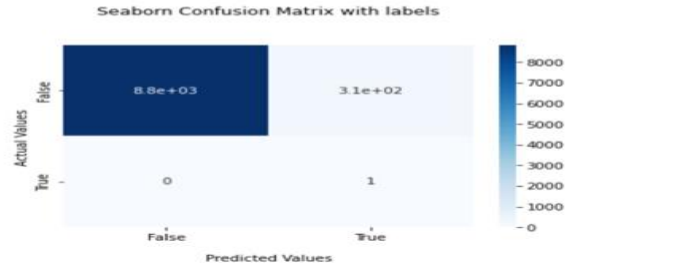


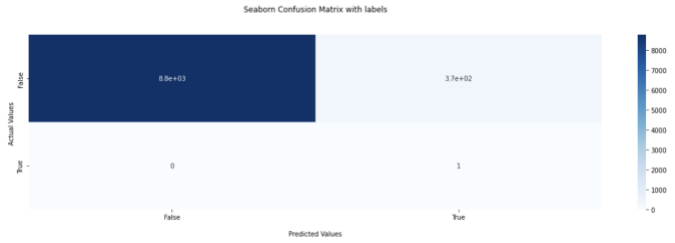
Fig : shows the evaluation scores of Random Forest classifier

2. **Gradient Boosted Classifier:** The iterative machine learning method to solve the classification problem is known as gradient boosting. This technique is based on ensemble learning in which the model is trained in such a way that errors of the previous iteration are used. Gradient Boosting(GB) accounts for misclassified samples by fitting a new learner to the ensemble residual that is the difference between the target outputs and the current predictions of the ensemble. Gradient Boosting tries to maximize the predictive power of the ensemble, i.e., minimize the bias. The advantage of using a boosting approach is generally high predictive power, but it comes with the cost of being slow to train as each new learner is trained sequentially. The GB model brought down the count of false positive from 371 to 0. Which is a very impressive figure.

```
[[8775  371]
 [   0    1]]
0.9594402536350716
              precision    recall  f1-score   support

           0       1.00      0.96      0.98      9146
           1       0.00      1.00      0.01         1

    accuracy                           0.96      9147
   macro avg       0.50      0.98      0.49      9147
weighted avg       1.00      0.96      0.98      9147
```

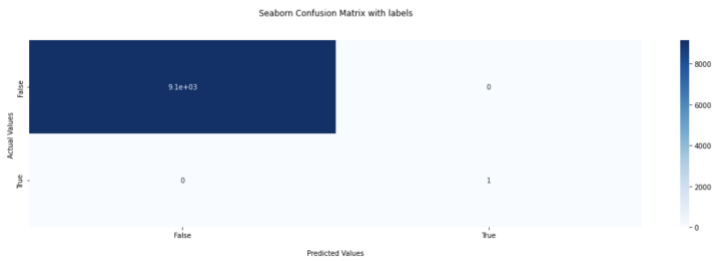


Above Fig shows evaluation before grid search.

```
[[9146    0]
 [   0    1]]
1.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      9146
           1       1.00      1.00      1.00         1

    accuracy                           1.00      9147
   macro avg       1.00      1.00      1.00      9147
weighted avg       1.00      1.00      1.00      9147
```



Above fig is evaluation of gradientboosted model after gridsearch.

Evaluation

The aim of the project was to predict the credit card default or not. The following machine learning models were applied and the best one amongst is selected as the prediction algorithm. For evaluation accuracy score along with confusion and classification matrix are used. Which altogether returns accuracy, precision, recall and F1_score along with the number of True positive and negative predictions.

```
[[9146    0]
 [   0    1]]
1.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      9146
           1       1.00      1.00      1.00         1

    accuracy                           1.00      9147
   macro avg       1.00      1.00      1.00      9147
weighted avg       1.00      1.00      1.00      9147
```

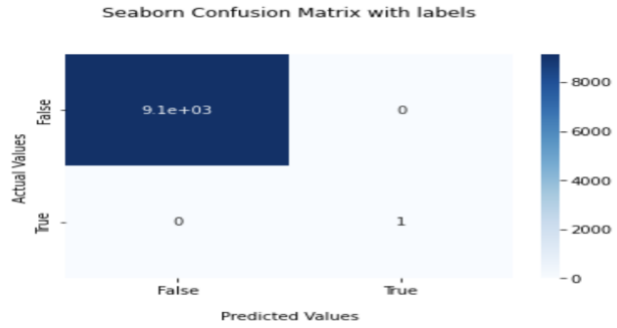


Fig : shows the evaluation scores and scatter plot of GB classifier

Selection of Model

Comparing the f1-score and and confusion matrix, gradient boosted decision tree performs better in avoiding false positive and negative predictions than random forest classifier.

Summary

Due to time constraint ,the assessment of algorithmic bias could not be done. However , it can be assessed with the help of AIF 360 python package which includes a comprehensive set of metrics for datasets and models to test for biases, explanations for these metrics, and

algorithms to mitigate bias in datasets and models.

A variety of algorithms can be used to mitigate bias. The choice of which to use depends on whether you want to fix the data (pre-process), the classifier (in-process), or the predictions (post-process). Learn more about how to choose.

○ **Reweighing**

Weights the examples in each (group, label) combination differently to ensure fairness before classification.

◉ **Optimized Pre-Processing**

Learns a probabilistic transformation that can modify the features and the labels in the training data.

○ **Adversarial Debiasing**

Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.

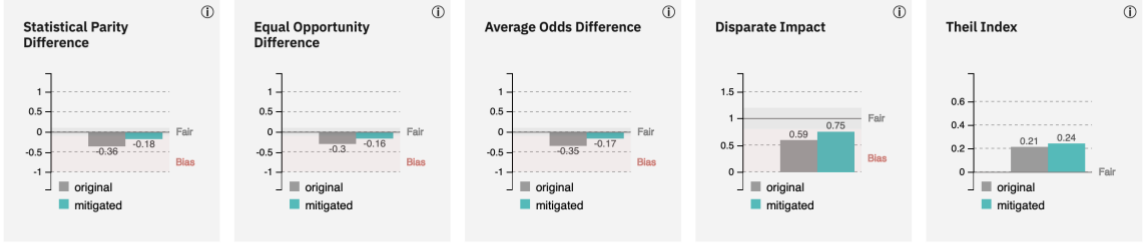
○ **Reject Option Based Classification**

Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

Privileged Group: ***Female***, Unprivileged Group: ***Male***

Accuracy after mitigation changed from 66% to 65%

Bias against unprivileged group unchanged after mitigation (4 of 5 metrics indicate bias)



Protected Attribute: Race

Privileged Group: ***Caucasian***, Unprivileged Group: ***Not Caucasian***

Accuracy after mitigation changed from 66% to 67%

Bias against unprivileged group was reduced to acceptable levels* for 4 of 4 previously biased metrics (0 of 5 metrics still indicate bias for unprivileged group)