**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

1. [Overview 5](#_Toc93483289)
2. [Investigation About Data 6](#_Toc93483290)
3. [Pre-processing 7](#_Toc93483291)
4. [Feature Selection 10](#_Toc93483292)
5. [Train Test split 11](#_Toc93483293)
6. [Modelling and Evaluation 11](#_Toc93483294)
7. [Task 1(Classification) 11](#_Toc93483295)
8. [Task 2(Regression) 14](#_Toc93483306)
9. [Selection of Model 16](#_Toc93483315)
10. [Prediction 17](#_Toc93483316)

# 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.

A picture containing text

Description automatically generated

1. 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 ‘.Chart, bar chart

   Description automatically generated
2. Analysing Numerical Variables and categorical variables:

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

Chart, histogram

Description automatically generated

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

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

* Marital status is analysed with help of pie chart.

Chart, pie chart

Description automatically generated

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

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

* 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.

Chart, box and whisker chart

Description automatically generated

* 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.

Chart

Description automatically generated

* Chart, line chart

  Description automatically generated

1. 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.
2. Outliers:

* Out of the continuous features age has few outliers .

Chart, box and whisker chart

Description automatically generated

Boxplot of outliers in continuous features

1. 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.
* Chart, treemap chart

  Description automatically generated

**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

# 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.

# 

# Fig : shows the evaluation scores of Random Forest classifier

**Gradient Boosted Classifier** : BOOSTING

# 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.

# Graphical user interface, application Description automatically generated

# Above Fig shows evaluation before grid search.

# Graphical user interface, application Description automatically generated

# 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.

# 

# 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. Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated