***Case Study – Credit Worthiness Classification***

[1]. Data was downloaded from [here](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)). File ‘german.data’ was used to retrieve the dataset and the data dictionary from the site was used to make sense out of the encoded data.

[2]. Data was splitted into train test set using *train\_test\_split* from *sklearn.model\_selection* library.

Also, the *stratify* parameter was set to response column for equal distribution of the response values in both the set.

[3]. Different hypothesis was tested and plotted to make inference on them. Also, statistical tests were performed on the features to make any conclusion. Data related to those can be found in the ‘CreditWorthinessClassification\_EDA.ipynb’ file.

[4]. Number of existing credit doesn’t seem to determine the good/bad behavior of the customer.

Test for column :NUM\_EXST\_CREDIT

-------------------------------------------------

contingency\_table :-

target 1 2

num\_exst\_credit

1 433 200

2 241 92

3 22 6

4 4 2

Observed Values :-

[[433 200]

[241 92]

[ 22 6]

[ 4 2]]

Expected Values :-

[[443.1 189.9]

[233.1 99.9]

[ 19.6 8.4]

[ 4.2 1.8]]

Degree of Freedom:- 3

chi-square statistic:- 2.6711981349760645

critical\_value: 7.814727903251179

p-value: 0.4451440800083001

Significance level: 0.05

Degree of Freedom: 3

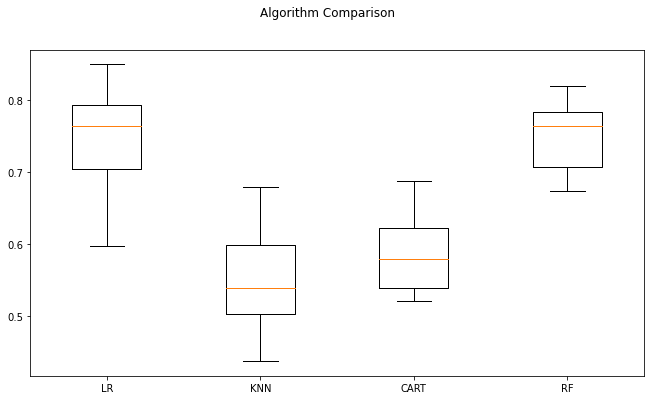
chi-square statistic: 2.6711981349760645

critical\_value: 7.814727903251179

Retain H0,There is no relationship between 2 categorical variables

Accept null hypothesis (Ho) where its states the number of existing credit is not useful in determining the good/bad behaviour of customer.

[5]-[7]. ‘CreditWorthinessClassification\_Model.ipynb’ contains the solution tested on different models such as Logistic Regression, K Neighbors Classifier, Decision Tree Classifier and Random Forest Classifier. Deta-ils on their performance can be found in this file. Note: **y-axis** *is roc\_auc\_score* value.



#### Since Random Forest model has the best performance with minimum std. deviation on multiple K-folds, and its ability to remain in-sensitive to outliers. It is chosen for building the prediction model.

[8]. Top three features according to the selected model importance.

|  |  |  |
| --- | --- | --- |
| *rank* | *feature* | *importance* |
| 0 | status\_of\_exst\_checking\_acnt\_A11 | 0.135029 |
| 1 | credit\_amnt | 0.131524 |
| 2 | duration\_mth | 0.122154 |

[9]. Based on the analysis done, provide a description of the “best” credit-worthy person. **Use language which will be suitable for stakeholder communication**.

A best credit worthy person can be said to have a

‘low credit amount’,

someone who is in their 30’s,

has his ‘own’ housing,

with saving account having at least 500 DM’s in it,

whose purpose is not to get a credit for used car, business, education and others (one of the categories).

who have none guarantors/debtors,

and have some kind of property with none other installment plans.

[10]. Document your solution as:

1. Solution developed (code): Refer ‘CreditWorthinessClassification\_EDA.ipynb’ and ‘CreditWorthinessClassification\_Model.ipynb’ file.
2. Process diagram (explaining all the steps you took to arrive at the solution)

Preprocessed the data for the input to the chosen models

Connected data with data dictionary to understand the encoded values and did EDA along with statistical test to remove the insignificant feature.

Train test split for model fitting and evaluation

Imported data from the given link

Model evaluation on the test set. Performance determined using several metrics

Choosing the model and tuning hyper-parameters using grid search

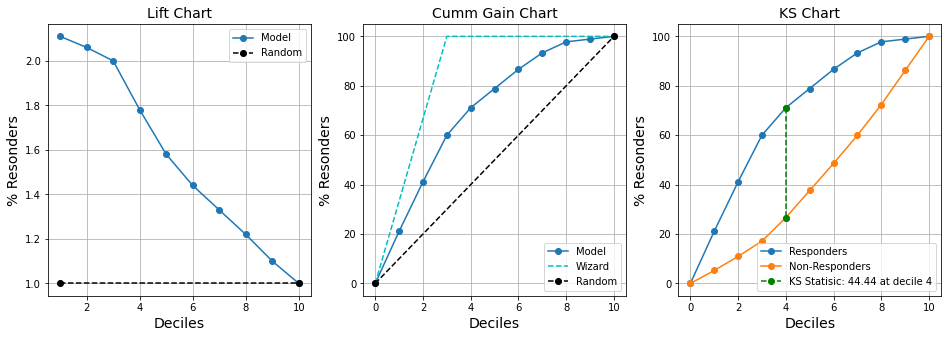
Model selection and performance

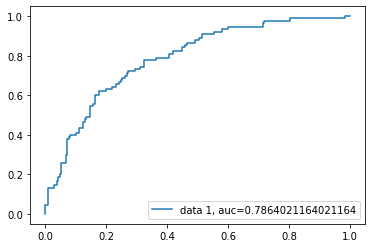
1. Explain model performance

Model performance is judged using the f1 score, recall, KS decile coverage and roc\_auc score. Accuracy in this case is not used since this is an imbalanced dataset.

1. Visualization to support your solution.

Below is the Lift, Cumulative gain chart and KS chart to support the effectiveness of model in determining the credit worthy person from a general population. This result is obtained on 30% test data.





**ROC\_AUC Curve**