SLDM ASSIGNMENT 2

GIRISH DESHPANDE (PRN: 21060641017)
SUDHANSHU DANDRIYAL (PRN: 21060641050)

BUSINESS PROBLEM

PERSONAL LOAN OFFER COMPANY WANTS TO IDENTIFY SOME SPECIFIC ATTRIBUTES THAT NEED TO BE CONSIDERED IN ORDER TO SIMPLIFY THE PROCESS OF SANCTIONING OF LOANS

METHODOLOGY

1. A TARGET VARIABLE WAS CREATED FOR CLASSIFYING THE LOANS AS 'GOOD' OR 'BAD'
2. RELEVANT VARIABLES WERE EXTRACTED FROM THE DATASET
3. IMPUTATION METHOD WAS USED FOR REPLACING THE MISSING VALUES IN THE VARIABLES THAT WERE CONSIDERED
4. OUTLIERS WERE IDENTIFIED AND WERE DROPPED FROM THE DATASET

METHODOLOGY

5 NON-NUMERICAL LABELS WERE CONVERTED TO NUMERICAL LABELS

6 ALL FEATURES WERE SCALED TO A RANGE OF 0 AND 1

7 LOGISTIC REGRESSION, RANDOM FOREST AND XGBOOST MODELS WERE USED FOR CLASSIFICATION OF LOANS

8 ALL THREE MODELS WERE COMPARED AND ACCURACY WAS CHECKED

VARIABLES CONSIDERED

- FROM THE CREDIT APPROVAL DATASET, 43 VARIABLES SUCH AS 'ID', 'MEMBER_ID', 'ZIP_CODE' ETC WERE DROPPED
- ONLY RELEVANT FEATURES LIKE 'LOAN_AMNT', 'TERM', 'INSTALLMENT' ETC WERE TAKEN INTO CONSIDERATION
- RELEVANT FEATURES WERE IDENTIFIED BASED ON THEIR SIGNIFICANCE DURING THE PROCESS OF LOAN APPROVALS

DEALING WITH MISSING VALUES

- MEAN IMPUTATION WAS CARRIED OUT FOR REPLACING THE MISSING VALUES IN NUMERICAL VARIABLES CONSIDERED
- WHEREAS, FOR NON-NUMERIC VARIABLES, THE MISSING VALUES WERE REPLACED WITH THE MOST FREQUENT VALUES OBSERVED

MODELS CONSIDERED FOR PREDICTION

- LOGISTIC REGRESSION
- RANDOM FOREST CLASSIFIER
- XGBOOST CLASSIFIER

LOGISTIC REGRESSION

- LOGISTIC REGRESSION WAS USED FOR PREDICTION OF LOANS (GOOD OR BAD) BY ANALYSING THE RELATIONSHIP AMONG THE VARIABLES CHOSEN
- CONFUSION MATRIX WAS USED FOR VISUALIZING THE PERFORMANCE OF THE MODEL
- **CONFUSION MATRIX:**

```
[[   0  253]
 [   3 1665]]
```

- THE F1 SCORE (HARMONIC MEAN OF THE PRECISION AND RECALL) WAS COMPUTED FOR CHECKING THE ACCURACY OF THE MODEL

RANDOM FOREST CLASSIFIER

- RANDOM FOREST CLASSIFIER WAS USED FOR PREDICTION OF LOANS BY USING A GROUP OF DECISION TREES
- CONFUSION MATRIX WAS USED FOR VISUALIZING THE PERFORMANCE OF THE MODEL
- **CONFUSION MATRIX:**

```
[[   5  248]
 [   6 1662]]
```

- THE F1 SCORE (HARMONIC MEAN OF THE PRECISION AND RECALL) WAS COMPUTED FOR CHECKING THE ACCURACY OF THE MODEL

XGBOOST CLASSIFIER

- XGBOOST WAS USED FOR CLASSIFICATION OF LOANS BY IMPLEMENTING GRADIENT BOOSTED DECISION TREES
- CONFUSION MATRIX WAS USED FOR VISUALIZING THE PERFORMANCE OF THE MODEL
- **CONFUSION MATRIX:**

```
[[  11  242]
 [  26 1642]]
```

- THE F1 SCORE (HARMONIC MEAN OF THE PRECISION AND RECALL) WAS COMPUTED FOR CHECKING THE ACCURACY OF THE MODEL

ACCURACY OF THE MODELS

F1 SCORE FOR LOGISTIC REGRESSION MODEL

0.928611266034579

F1 SCORE FOR RANDOM FOREST CLASSIFIER MODEL

0.929010620458356 6

F1 SCORE FOR XGBOOST CLASSIFIER MODEL

0.924549549549549 5

BASED ON THE F1 SCORES OBTAINED FOR ALL THE MODELS, WE CAN CONCLUDE THAT RANDOM FOREST CLASSIFIER IS THE MOST ACCURATE ONE FOR PREDICTING WHICH LOANS CAN BE CLASSIFIED AS GOOD OR BAD

CONCLUSION

- SINCE THE MAIN OBJECTIVE WAS TO CLASSIFY THE LOANS AS 'GOOD' OR 'BAD' BASED ON THE VARIOUS ATTRIBUTES, CLASSIFICATION MODELS SUCH AS LOGISTIC REGRESSION, RANDOM FOREST AND XGBOOST WERE TAKEN INTO CONSIDERATION
- BASED ON THE F1 SCORES FOR ALL THE MODELS, RANDOM FOREST CLASSIFIER WAS IDENTIFIED AS THE MOST ACCURATE MODEL FOR THE PURPOSE OF SANCTIONING OF THE LOANS

THANK YOU