LOAN

ANALYSIS

Brief about DATASET

- It contains 25 columns and 30,000 rows.
- There are no NA values in the dataset.
- Default Payment is our dependent variable, which consist of 2 variables 1 and 0, which describes s/he will pay the loan and s/he will not pay the loan.
- The rest columns contain their data like previous payments, gender, qualification, and their credit amount.

STEPS

- CHANGING THE COLUMNS TO FACTOR, WHICH CONTAINS FACTOR ELEMENTS.
- CHEKING THE BIASNESS IN THE DEPENDENT VARIABLE.
- CHANGING GENDER COLUMN ELEMENTS TO THE RELEVANT VALUE

  1 STANDS FOR MALE AND 2 STANDS FOR FEMALE.
- CHANGING QUALIFICATION COLUMN ELEMENT TO RELEVANT VALUES AS GIVEN

  IN THE DESCRIPTION FILE.
- VISUALIZING THE DATA.
- RANDOMLY SHUFFLING DATA.
- REMOVING INSIGNIFICANT COLUMNS
- SUBESTTING THE DATASET INTO TEST AND TRAIN DATASET.
- APPLYING THE MACHINE LEARNING ALGORITHM TO FIND THE BEST MODEL.

Changing columns to factor

"Gender",

"Academic_Qualification",

"Marital",

"Repayment_Status_Jan“

"Repayment_Status_Feb",

"Repayment_Status_March",

"Repayment_Status_April",

"Repayment_Status_May",

"Repayment_Status_June",

"Default_Payment"

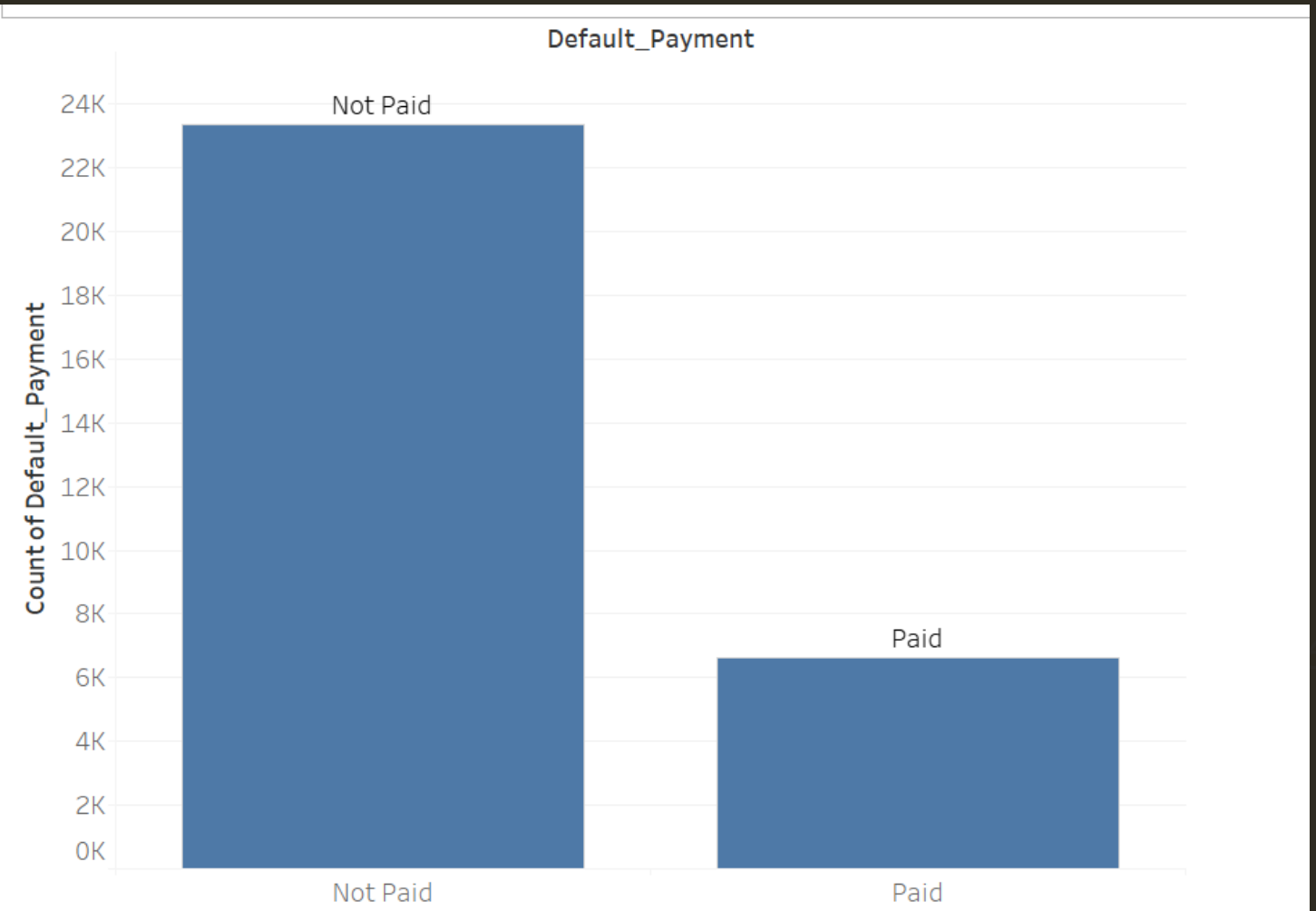
These are columns that are needed to be changed

Changing elements of columns

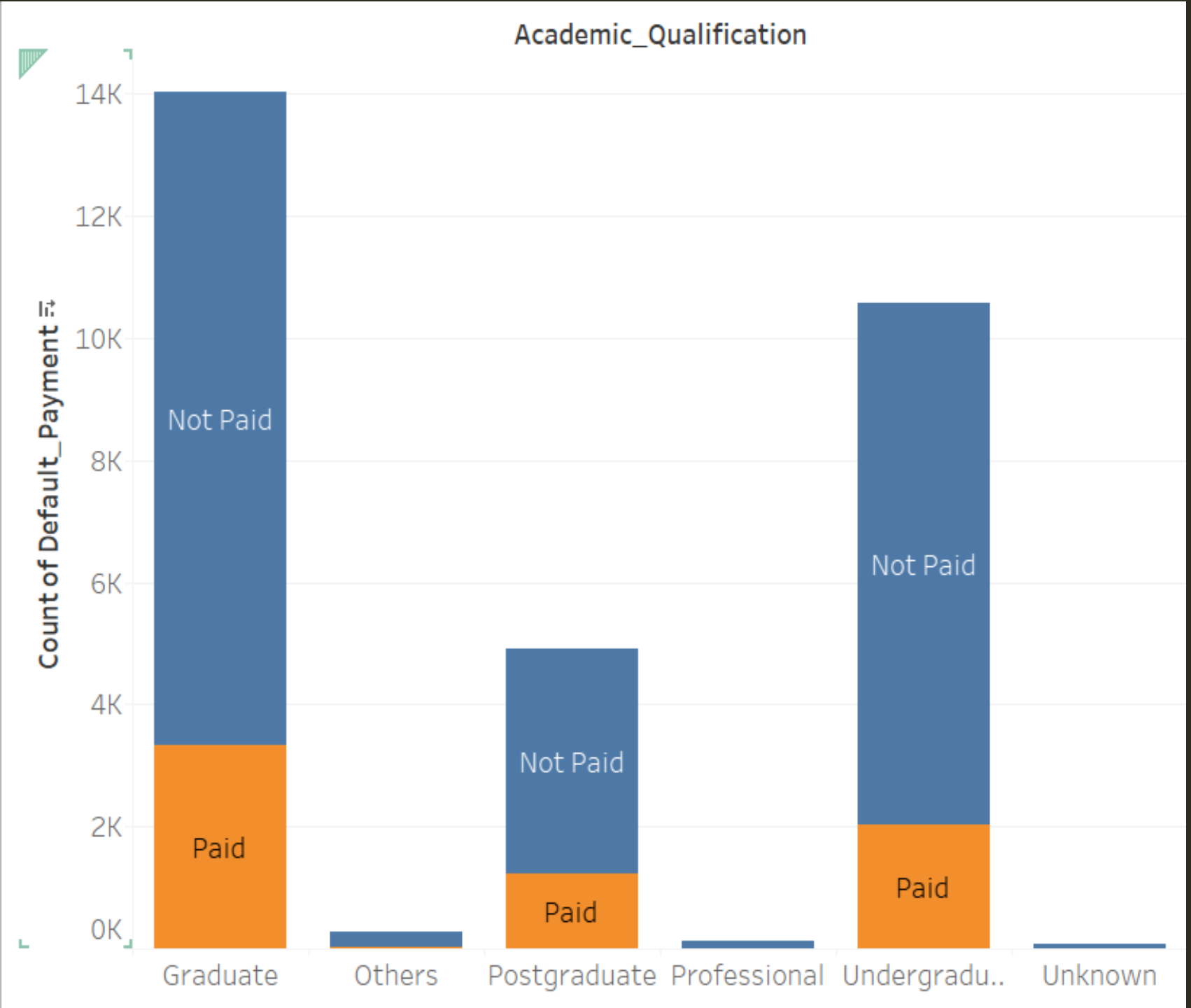
IN GENDER COLUMN:
1=MALE
2=FEMALE

IN QUALIFICATION COLUMN:
1=UNDERGRADUATE
2=GRADUATE
3=POSTGRADUATE
4=PROFESSIONAL
5=OTHER
6=UKNOWN

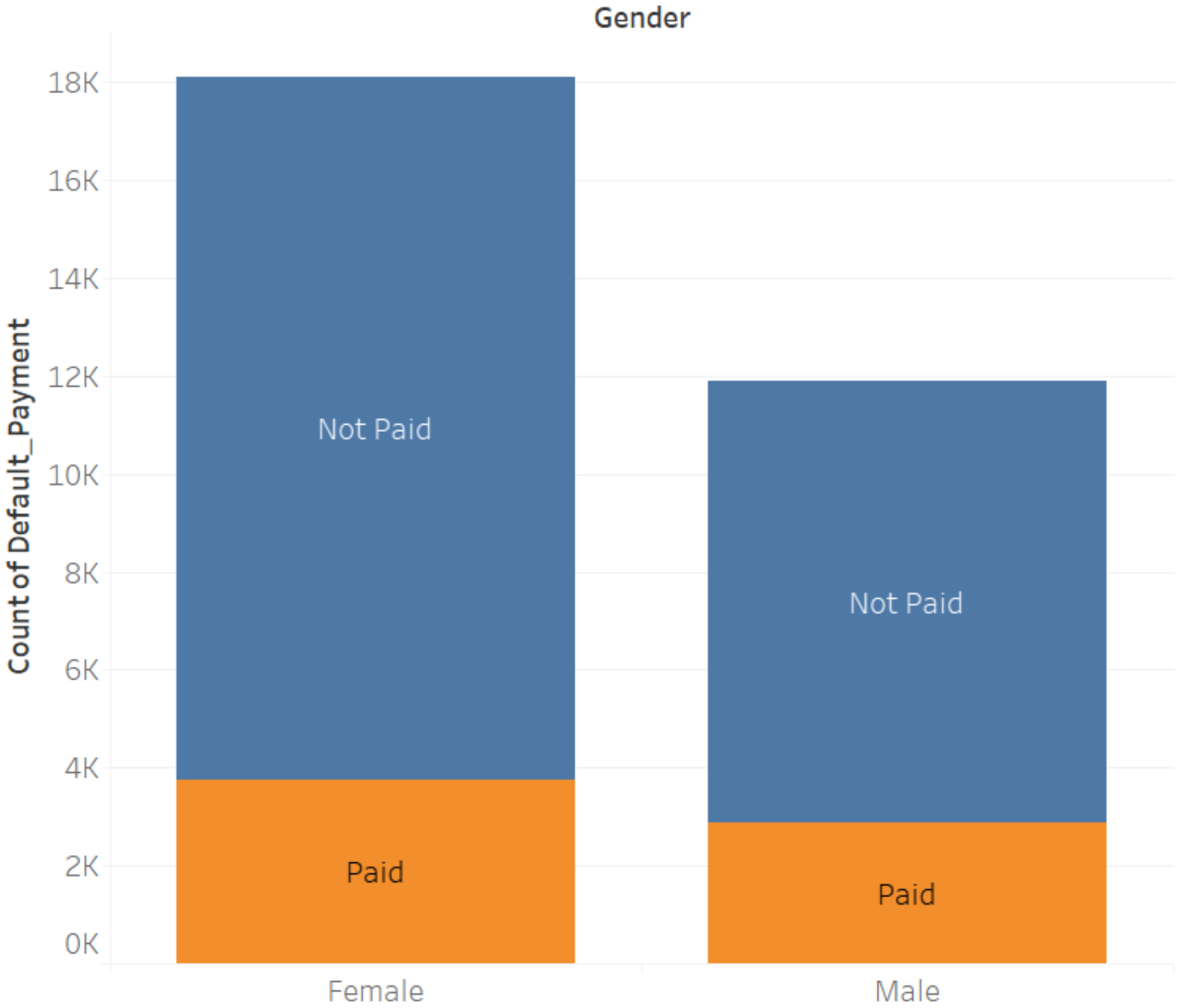
CHECKING THE BIASNESS



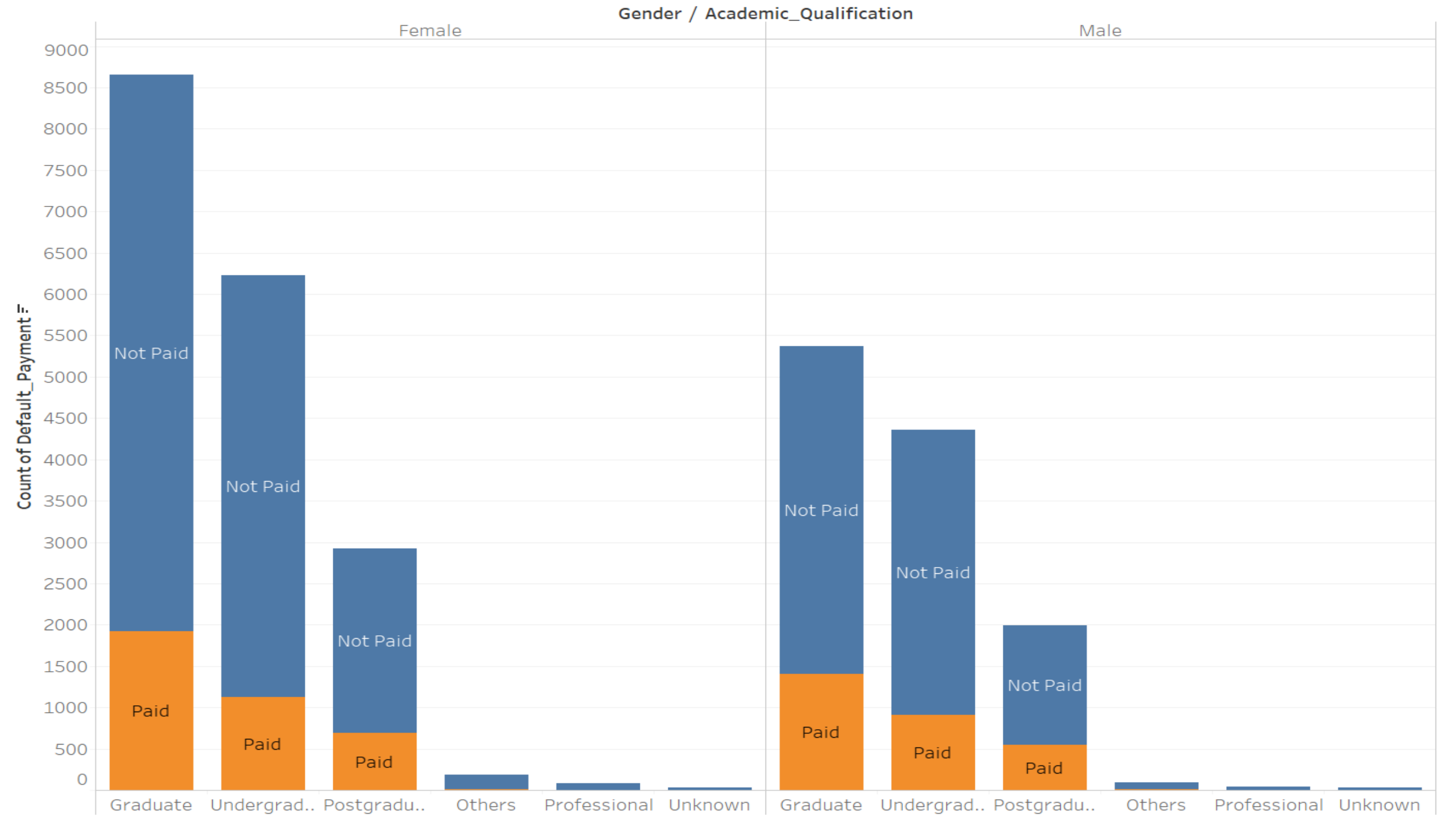
0 (Not Paid) IS OCCURING 23364 TIMES IN THE DATASET WHILE 1(Paid) IS OCCURING ONLY 6636 TIMES, CLEARLY THERE IS CLASS BIASNESS IN THE MODEL.



MOSTLY GRADUATE, POSTGRADUATE AND UNDERGRADUATE PERSONS FAILED TO PAY UP THE LOAN.



CLEARLY THERE ARE MORE FEMALES WHO FAILED TO PAY THE LOAN.


Gender / Academic_Qualification
Female
Male
Count of Default_Payment
9000
8500
8000
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Not Paid
Paid
Not Paid
Paid
Not Paid
Paid
Not Paid
Paid
Not Paid
Paid
Not Paid
Paid
Graduate
Undergrad..
Postgradu..
Others
Professional
Unknown
Graduate
Undergrad..
Postgradu..
Others
Professional
Unknown

RANDOMLY SHUFFLING DATA TO AVOID BIASNESS

Sub setting into training and test data set

Training data contains 70% of actual dataset.

Test data contains the 30% of actual dataset.

APPLYING ALGORITHMS

USING LOGISTIC REGRESSION:

ACCURACY OF THE MODEL: 0.8255

USING SVM:

ACCURACY OF THE MODEL:0.8375