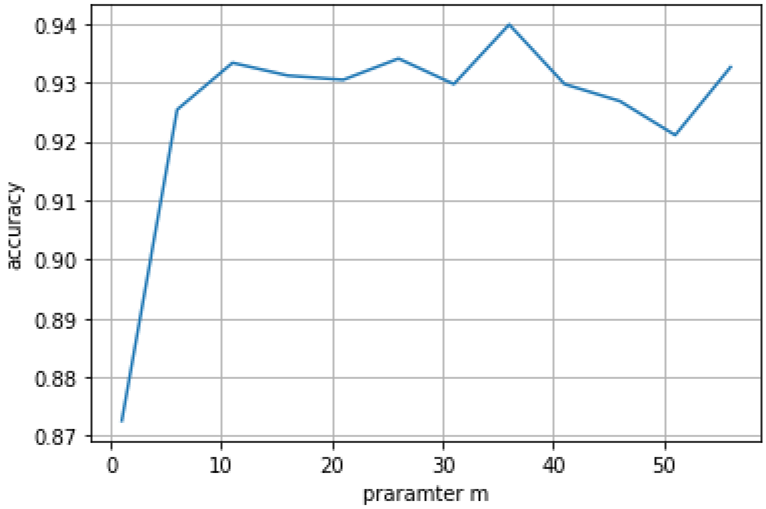
# Assignment-3

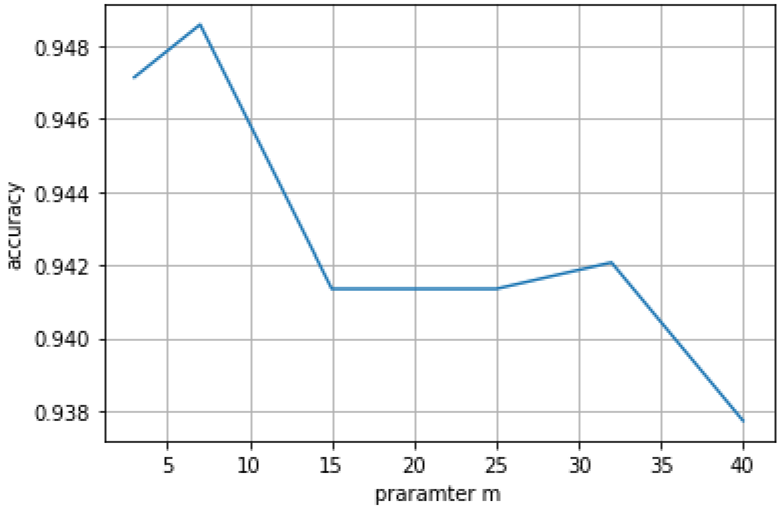
AI20BTECH11006

Question 4)

a) The code can be found in q4.ipynb
Sklearn's inbuilt random forest is more efficient than the random forest written by me, it gives higher accuracy and takes much less time

b) Two plots of accuracy can be found below





As can be seen from the plot, there isn't much consistency in which value of m gives a better accuracy, but in general, sqrt(M) seems to give a decent accuracy value compared to other values.

c) The results aren’t all that consistent, a set of values I recorded is mentioned below

```
m: 7
The oob error is : 0.0670807453416149
m: 14
The oob error is : 0.0661490683229814
m: 21
The oob error is : 0.07329192546583851
m: 35
The oob error is : 0.07391304347826089
These values were recorded for 10 trees.
```

Question 5)

a) I removed all the attributes which had a most/all values as NA, other columns such as desc, url which seemed of no use were also removed.
I removed the rows which had NA in some column or other after doing the above, the reason to do this is that we have a large dataset, removing some part of dataset wouldn’t cause a problem.
Here is a list of columns I removed

```
'id','member_id','url','desc','emp_title','purpose','title','zip_cod
e','addr_state','earliest_cr_line'
'inq_last_6mths','sub_grade','application_type','last_pymnt_d','issu
e_d','pymnt_plan'
```

I removed sub grade because of the dependency between interest rate and grades, addr_state also seems like a proxy variable.

b) The maximum accuracy is achieved for
n_estimators = 100, max_features = ’auto’, max_depth=7
The depth doesn’t matter after 7,
The number of trees (n_estimators) seems to play a role, but the change in accuracy is within 0.1%,

Testing against a simple decision tree, I obtained the following values

```
The accuracy of the inbuilt decision tree classifier is:
0.9916279069767442
The precision of the inbuilt decision tree classifier is:
0.9951133204145253
The recall of the inbuilt decision tree classifier is:
0.9950294860994103
The accuracy of the gradient boosting model is: 0.9967084078711985
The precision of the gradient boosting model is: 0.9964729593550554
The recall of the gradient boosting model is: 0.9996630160067397
```

Even decision tree performs rather good on this dataset, the boosting algorithm helps improve the accuracy by around 0.5%