**1. Exploratory Data Analysis:**

We got following datasets

1. Account
2. Enquiry
3. Customer Data

Following are the size of Data

Enquiry Train: (413188, 6)

Enquiry Test : (173499, 6)

Account Train: (186329, 21)

Account Test: (79568, 21)

Customer Data Train: (23896, 34)

Customer Data Test: (10240, 34)

**Enquiry Data:**

* The enquiry data has the total number of 6 columns where it has the key index of customer id.
* Each customer id has multiple enquiries and data is collected for each enquiry

**Account Data**

* The account data is based on transaction history which has 21 columns.
* The payment history is the main key for the rest of the feature where each customer has multiple payment history with payment date, credit limit, cash limit, current balance…etc.
* The payment history contains the values of Days Past Due for 18 months.
* We can see some correlations in the data
* Some variable has correlation between them as follows

Top Absolute Correlations for Account data

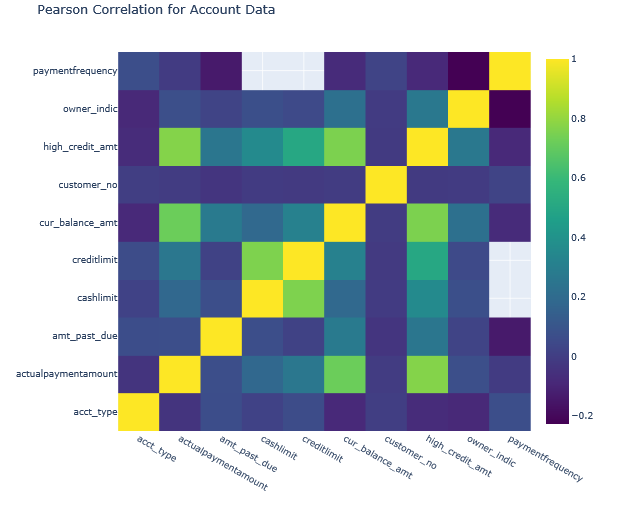
actualpaymentamount high\_credit\_amt 0.769984

cashlimit creditlimit 0.758459

cur\_balance\_amt high\_credit\_amt 0.755600

actualpaymentamount cur\_balance\_amt 0.721505

creditlimit high\_credit\_amt 0.502603



**Customer Data**

* The customer data segment has the large number features based on customer bank account.
* The target variable is 'Bad\_label' and it is based on customer id as key index.
* The target variable is Customer creditworthiness either with positive or negative flag.
* We can find that there are some columns which are very correlated with each other.
* I have written a function to get the top values

Top Absolute Correlations for Data file

feature\_34 feature\_68 1.000000

feature\_35 feature\_69 1.000000

feature\_34 feature\_39 0.995821

feature\_39 feature\_68 0.995821

feature\_29 feature\_44 0.983200

feature\_19 feature\_55 0.918355

feature\_56 feature\_64 0.765519

feature\_39 feature\_41 0.763549

* The overall customers falling under the negative bad label is 4.29% which means 95.71% of the customers having the positive creditworthiness.
* From the correlation heat map, the feature 35 and 69 has the correlation value of 1 which means that the features are same and repeated features. The feature 34 and 68 have the same repeated feature which has the correlation value of 1.
* The feature 29, 66 and 44 has second highest correlation between them where these two features are representing the area postcode of the customer address or bank address.
* The feature 3 and 7 has the third highest correlation between them.
* The actual payment amount and high credit amount has the highest correlation of 0.769 in account segment
* The cash limit and credit limit has the correlation value of 0.758
* The current balance amount and high credit amount has the next highest correlation value of 0.755

Following is the table of missing values percentages more than 40 %

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Total Missing Values** | **Missing Percentage** |
| **feature\_61** | 34121 | 0.999561 |
| **feature\_74** | 34112 | 0.999297 |
| **feature\_18** | 34105 | 0.999092 |
| **feature\_10** | 34065 | 0.997920 |
| **feature\_49** | 33977 | 0.995342 |
| **feature\_17** | 32681 | 0.957376 |
| **feature\_9** | 32343 | 0.947475 |
| **feature\_8** | 32343 | 0.947475 |
| **feature\_57** | 30705 | 0.899490 |
| **feature\_73** | 29917 | 0.876406 |
| **feature\_48** | 26365 | 0.772352 |
| **feature\_6** | 23905 | 0.700287 |
| **feature\_45** | 19618 | 0.574701 |
| **feature\_13** | 18522 | 0.542594 |
| **feature\_53** | 16594 | 0.486114 |
| **feature\_51** | 16338 | 0.478615 |

**2. Feature Engineering:**

We have created multiple features while we ran the algorithm

* mean\_diff\_lastpaymt\_opened\_dt
* total\_diff\_lastpaymt\_opened\_dt
* count\_enquiry\_recency\_365
* mean\_diff\_open\_enquiry\_dt
* payment\_history\_mean\_length
* avg\_enq\_amount
* Ratio\_currbalance\_creditlimit
* utilisation\_trend
* max\_freq\_enquiry

Following is the feature Ranking for Features with more than .005 Gain. We got this from RF classifier.

Feature ranking:

1. feature 26 (0.032800)

2. feature 8 (0.029422)

3. feature 32 (0.027268)

4. feature 33 (0.027267)

5. feature 35 (0.027035)

6. feature 34 (0.025493)

7. feature 2 (0.023953)

8. feature 30 (0.023834)

9. feature 3 (0.021718)

10. feature 17 (0.019917)

11. feature 5 (0.016883)

12. feature 8008 (0.014782)

13. feature 1 (0.014771)

14. feature 0 (0.012498)

15. feature 12 (0.010687)

16. feature 7999 (0.010076)

17. feature 31 (0.009880)

18. feature 14 (0.009714)

19. feature 21 (0.009275)

20. feature 7998 (0.009142)

21. feature 22 (0.009131)

22. feature 9 (0.008978)

23. feature 37 (0.008228)

24. feature 85 (0.008180)

25. feature 25 (0.007978)

26. feature 36 (0.007526)

27. feature 41 (0.007160)

28. feature 18 (0.006744)

29. feature 44 (0.006179)

30. feature 5660 (0.006061)

31. feature 934 (0.006040)

32. feature 15 (0.005911)

33. feature 69 (0.005696)

34. feature 730 (0.005432)

35. feature 107 (0.005421)

36. feature 77 (0.005094)

37. feature 919 (0.005003)

**3.** **Model Building**

We have used simple model of Random Forest with following parameters

rf\_params = {

'n\_jobs': -1,

'n\_estimators': 100,

'warm\_start': True,

'max\_features': 0.2,

'max\_depth': 9,

'min\_samples\_leaf': 2,

'max\_features' : 'sqrt',

'random\_state' : seed,

'verbose': 0

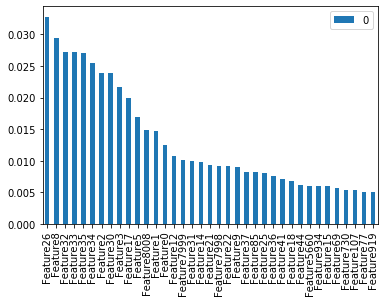
}

We can tune and use grid search in future.

By default the method is gini in random forest

**Model Accuracy = 0.9559051628628407**

**Feature Importance of top columns**

. . .