**Data Description**

In the given data set we have 41 columns.

* 1 Target Variable – loan\_default
* 40 Independent variables

**Numerical Attributes:** UniqueID,disbursed\_amount,asset\_cost,ltv,branch\_id,supplier\_id,manufacturer\_id,Current\_pincode\_ID,State\_ID,Employee\_code\_ID,MobileNo\_Avl\_Flag,Aadhar\_flag,PAN\_flag,VoterID\_flag,Driving\_flag,Passport\_flag,PERFORM\_CNS.SCORE,PRI.NO.OF.ACCTS,PRI.ACTIVE.ACCTS,PRI.OVERDUE.ACCTS,PRI.CURRENT.BALANCE,PRI.SANCTIONED.AMOUNT,PRI.DISBURSED.AMOUNT,SEC.NO.OF.ACCTS,SEC.ACTIVE.ACCTS,SEC.OVERDUE.ACCTS,SEC.CURRENT.BALANCE,SEC.SANCTIONED.AMOUNT,SEC.DISBURSED.AMOUNT,PRIMARY.INSTAL.AMT,SEC.INSTAL.AMT,NEW.ACCTS.IN.LAST.SIX.MONTHS,DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS, NO.OF\_INQUIRIES, loan\_default

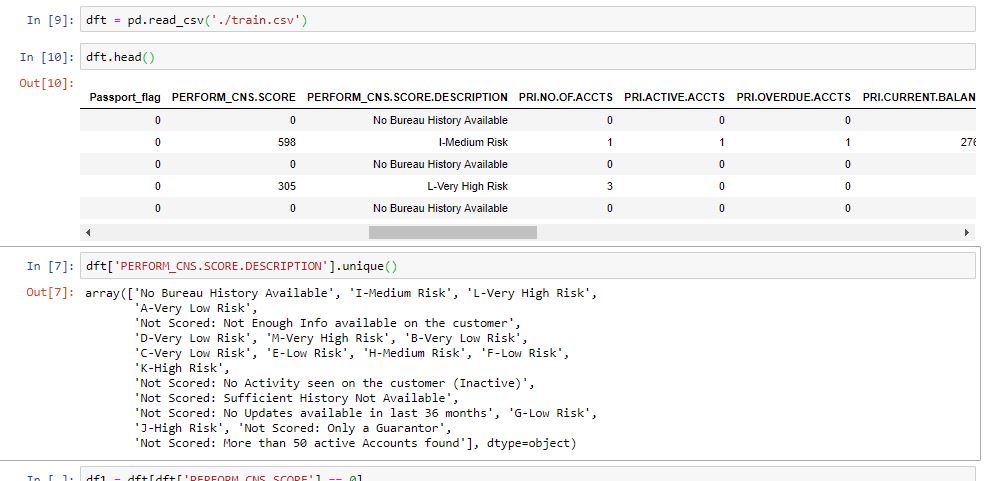
**Categorical Attributes:**

Employment.Type, Date.of.Birth, DisbursalDate, PERFORM\_CNS.SCORE.DESCRIPTION, AVERAGE.ACCT.AGE, CREDIT.HISTORY.LENGTH

**Exploratory Data Analysis and Data Preparation**

If we observe the data, customer data is divided into two sections

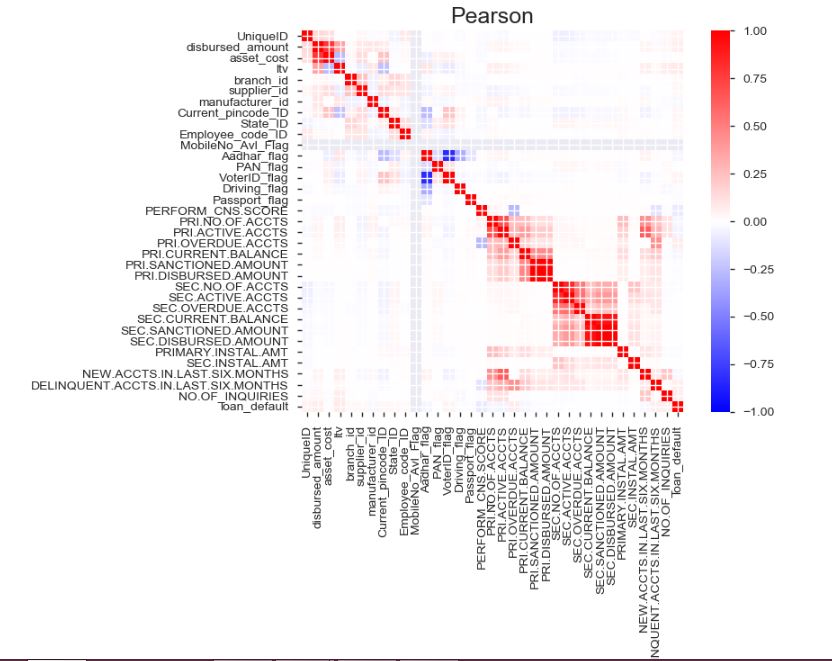
1. Customer with past historical data
2. Customer without any historical data(i.e new customer)



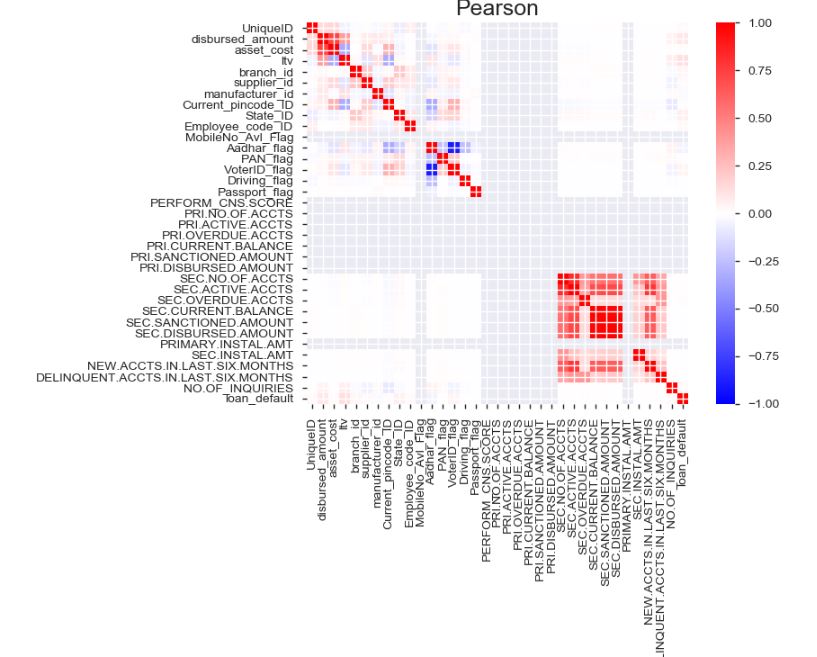
So Data is divided into two sections and each section is having their own model.

**Correlation b/w different attributes:**

1. **Data with past history:**

****

1. **Data without past history:**

****

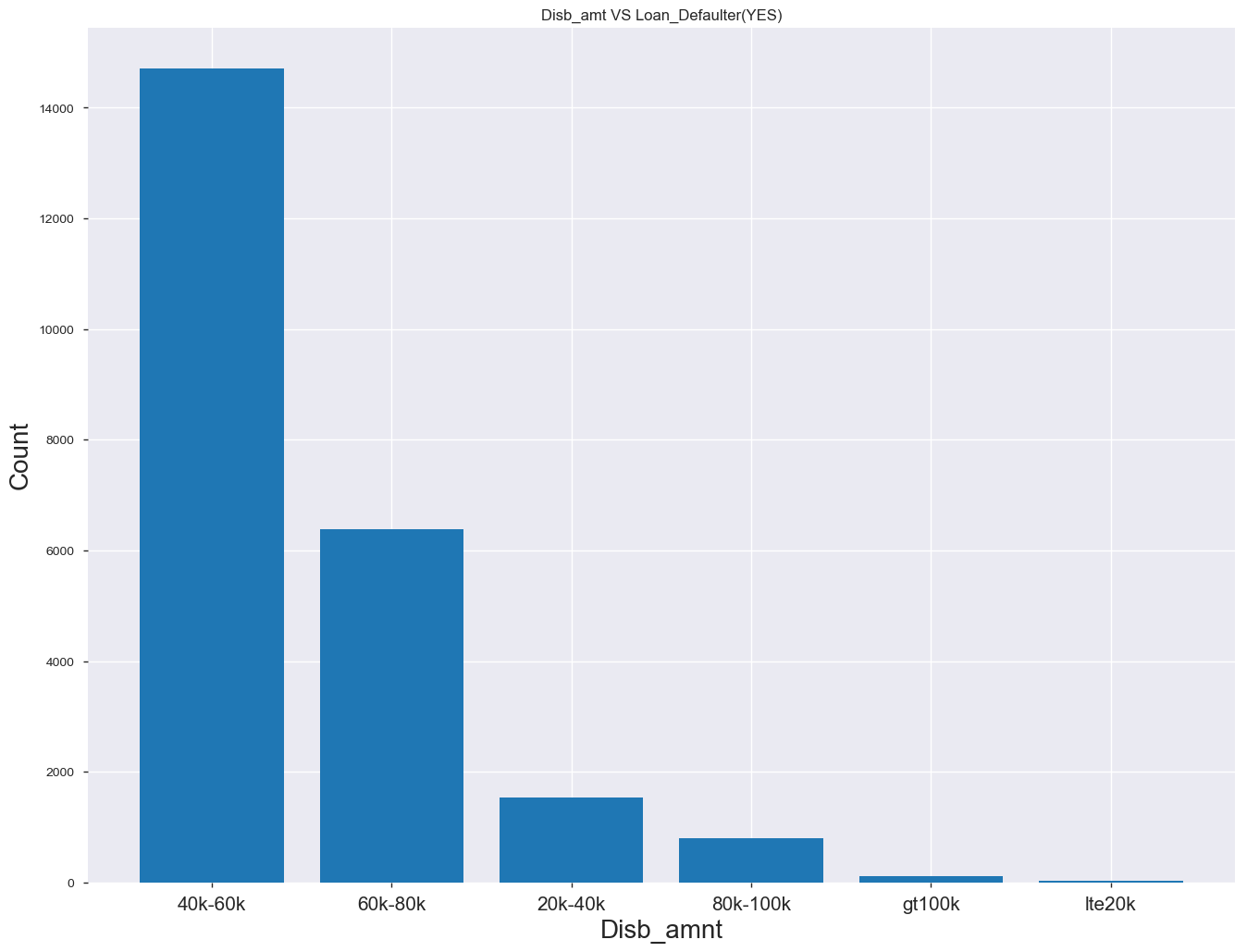
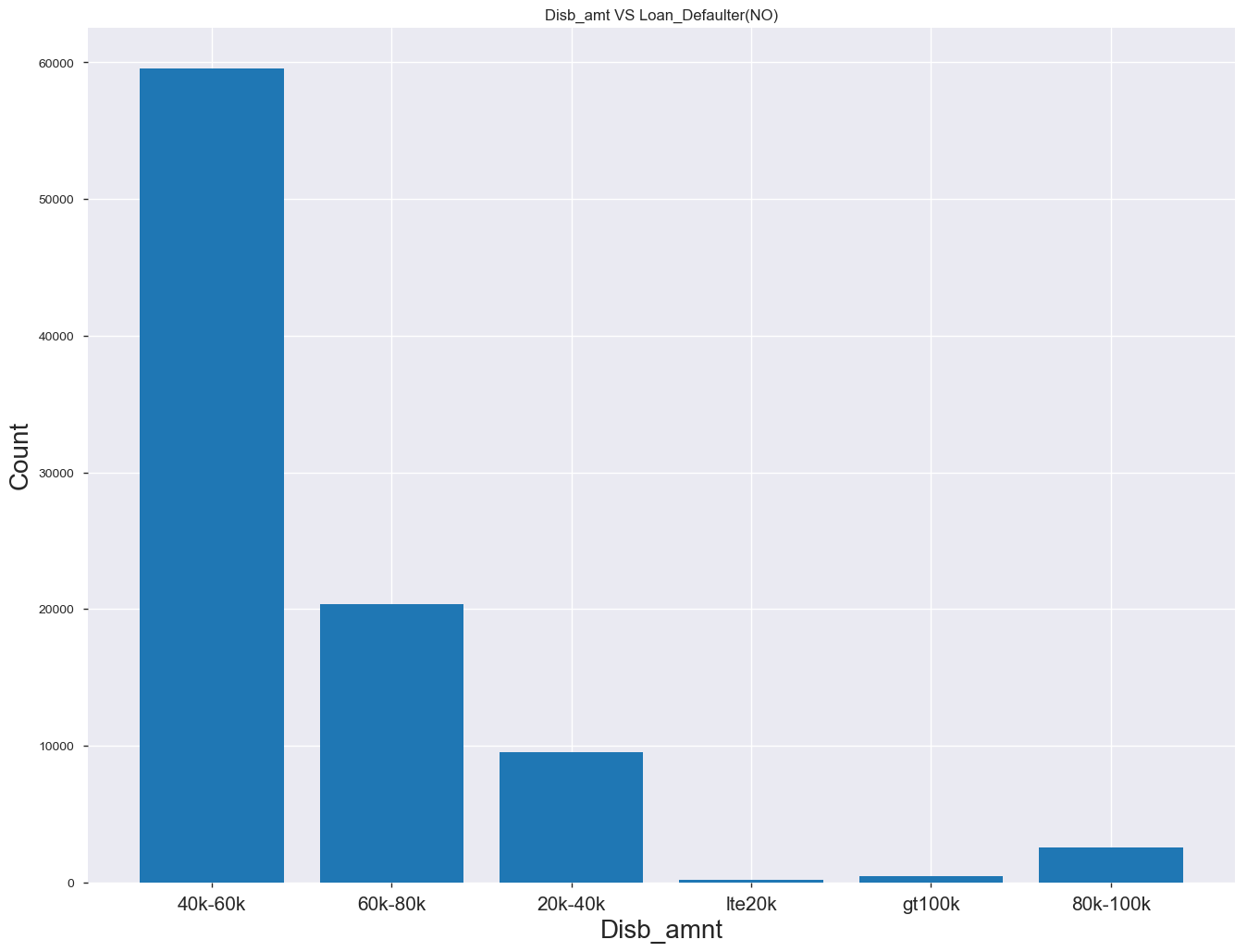
**Non–Significant features**:

1. **Data with past history:** PRI.DISBURSED.AMOUNT,SEC.DISBURSED.AMOUNT,SEC.SANCTIONED.AMOUNT,Mobile\_Avl\_Flag
2. **Data without past history:**

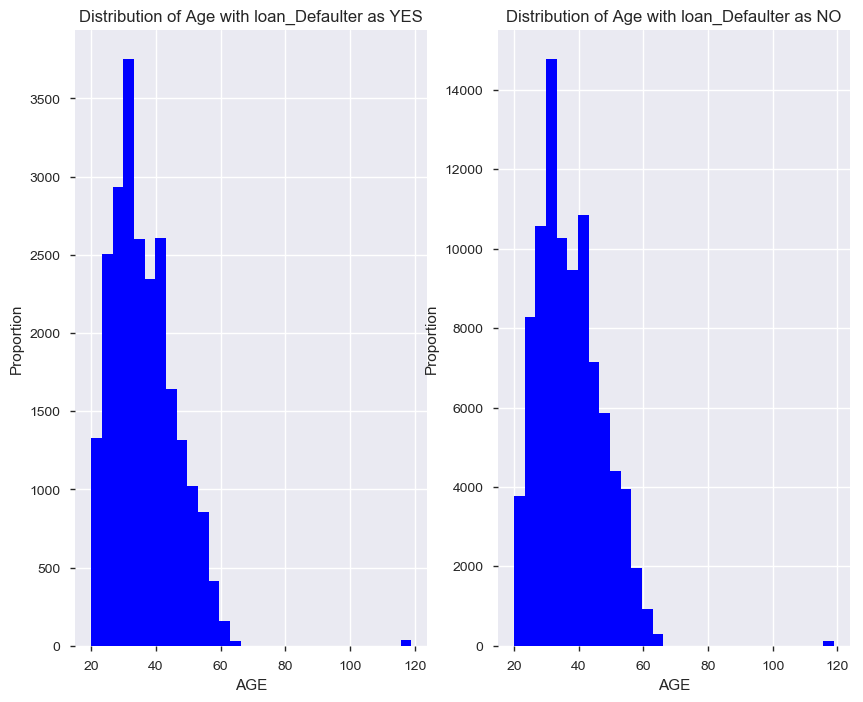
SEC.DISBURSED.AMOUNT,SEC.SANCTIONED.AMOUNT,MobileNo\_Avl\_Flag,PERFORM\_CNS.SCORE.DESCRIPTION,PERFORM\_CNS.SCORE,PRI.ACTIVE.ACCTS,PRI.CURRENT.BALANCE,PRI.DISBURSED.AMOUNT,PRI.NO.OF.ACCTS,PRI.OVERDUE.ACCTS,PRI.SANCTIONED.AMOUNT,PRIMARY.INSTAL.AMT,NEW.ACCTS.IN.LAST.SIX.MONTHS,DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS

**For Data with past history:**

**Distribution of Disbursed amount w.r.t loan\_Defaulter:**

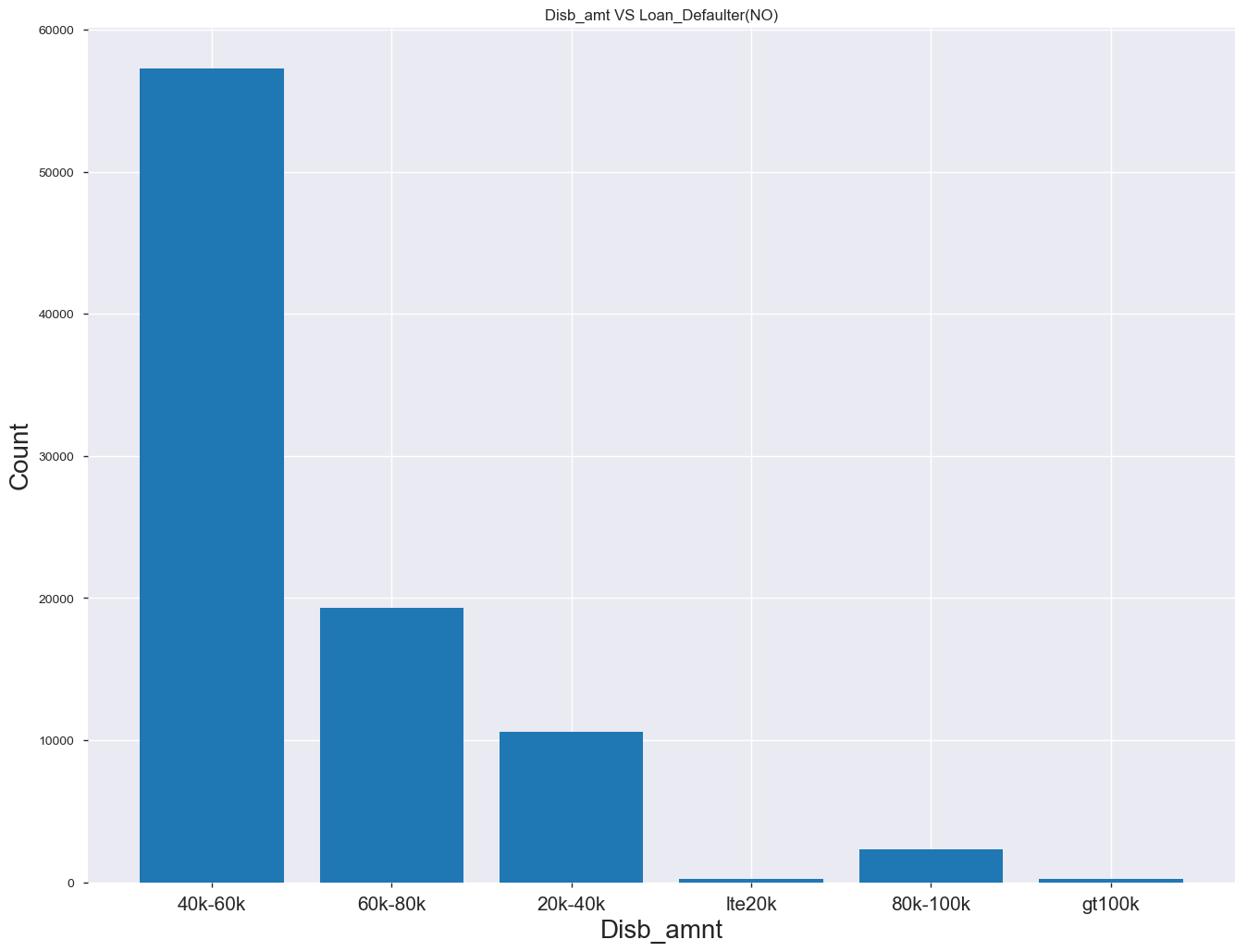
****

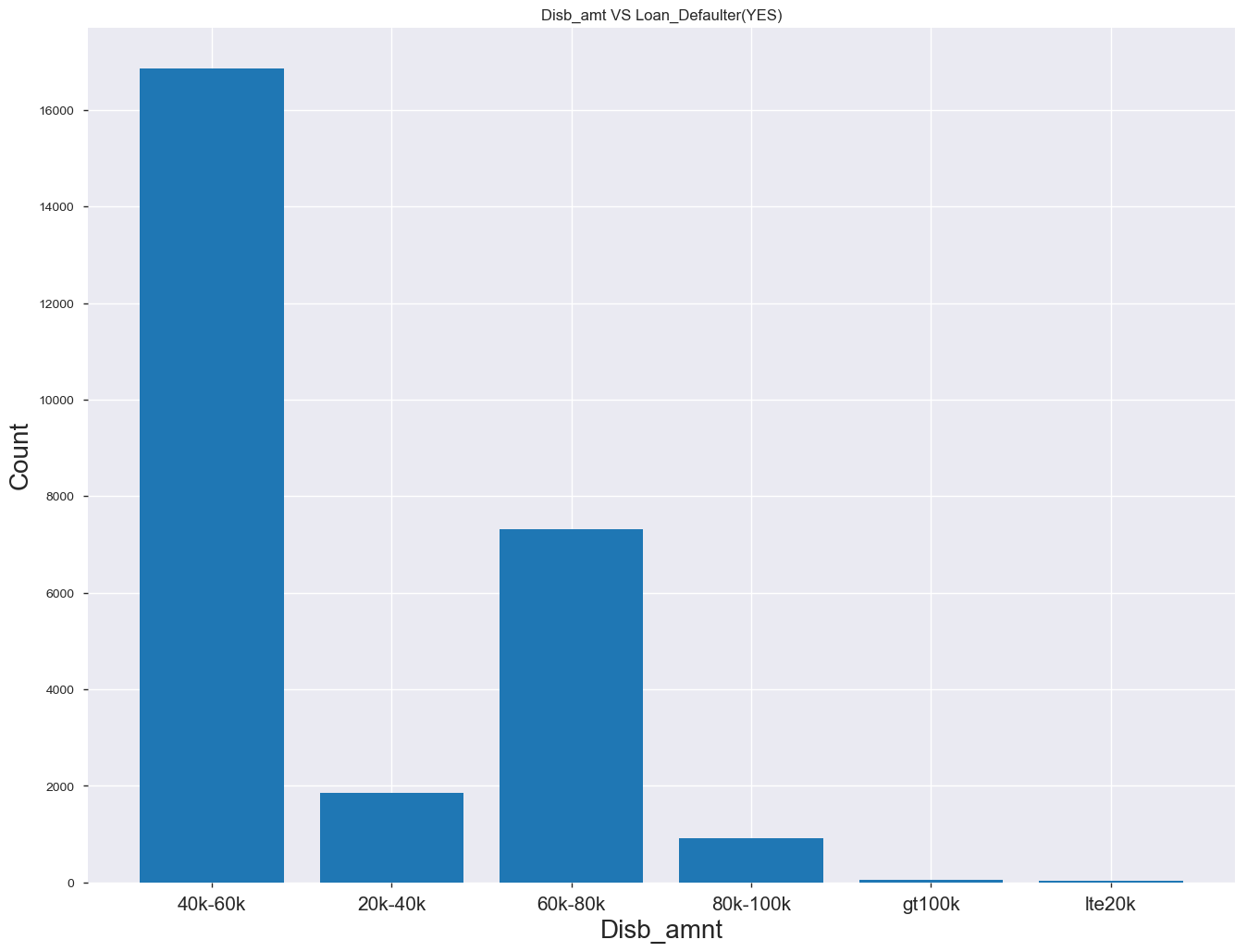
**Distribution of Age w.r.t loan\_default:**

****

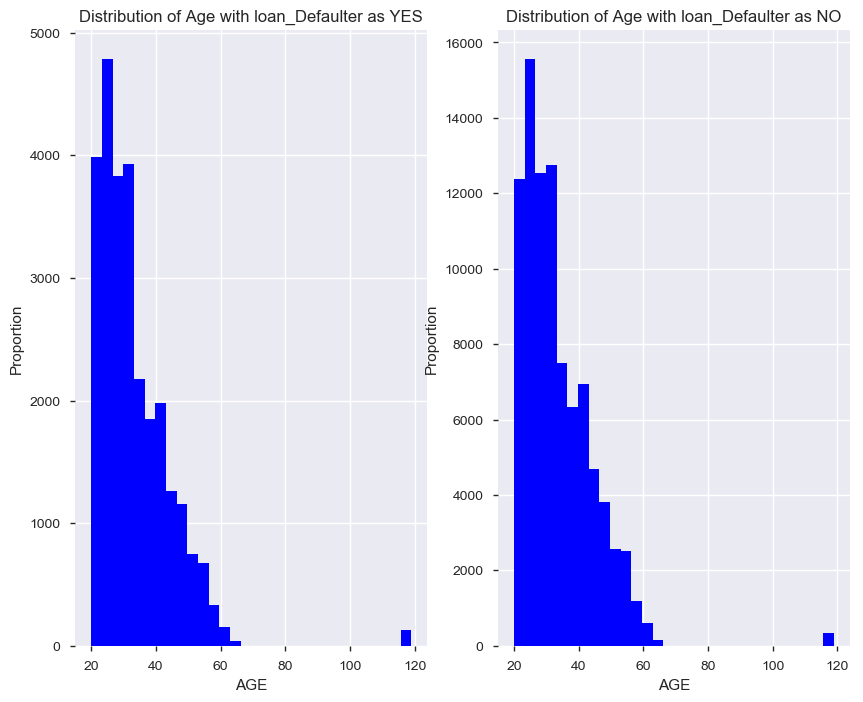
**For Data without past history:**

**Distribution of Disbursal Amount w.r.t loan\_default:**





**Distribution of Age w.r.t loan\_Default:**

****

**Feature Selection**

**For Data with past history:**

Employment.Type,PERFORM\_CNS.SCORE.DESCRIPTION,proof\_submitted,Age\_conv,Disb\_amnt\_conv,bidbidld\_ratio\_conv,sidsidld\_ratio\_conv,midmidld\_ratio\_conv,pinidpinld\_ratio\_conv,stidstateidld\_ratio\_conv,empidempidld\_ratio\_conv,PriPriOvrDue\_ld\_ratio\_conv,PriPriActAcnt\_ld\_ratio\_conv,PriPriPrevAcnt\_ld\_ratio\_conv,Pri\_Current\_Bal\_Conv,Pri\_Sanctioned\_Amt\_Conv,Pri\_Inst\_Amt\_Conv,SecSecOvrDue\_ld\_ratio\_conv,SecSecActAcnt\_ld\_ratio\_conv,SecSecPrevAcnt\_ld\_ratio\_conv,Sec\_Inst\_Amt\_Conv,Sec\_left\_bal\_conv,AvgAccountAge\_Conv,CreditHistAge\_Conv,PRI\_left\_bal\_Conv,ltv\_Conv,NewAcctInLast6Mnths,DelinquentAccntInLast6Mnths,No\_Inquiries,loan\_default

**For Data without past history:**

Employment.Type,proof\_submitted,Age\_conv,Disb\_amnt\_conv,bidbidld\_ratio\_conv,sidsidld\_ratio\_conv,midmidld\_ratio\_conv,pinidpinld\_ratio\_conv,stidstateidld\_ratio\_conv,empidempidld\_ratio\_conv,SecSecOvrDue\_ld\_ratio\_conv,SecSecActAcnt\_ld\_ratio\_conv,SecSecPrevAcnt\_ld\_ratio\_conv,Sec\_Inst\_Amt\_Conv,Sec\_left\_bal\_conv,AvgAccountAge\_Conv,CreditHistAge\_Conv,ltv\_Conv,No\_Inquiries,loan\_default

**Choosing the classification model:**

1. **Data With Past History:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AUC(ROC) | F1 | Precision | Recall |
| RandomForest | **0.93** | **0.93** | **0.935** | **0.935** |
| Logistic Regression | **0.60** | **0.60** | **0.60** | **0.60** |
| AdaBoostClassifier | **0.67** | **0.67** | **0.675** | **0.675** |
| GaussianNB | **0.62** | **0.62** | **0.62** | **0.63** |

1. **Data without past data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AUC(ROC) | F1 | Precision | Recall |
| RandomForest | **0.56** | **0.56** | **0.56** | **0.57** |
| Logistic Regression | **0.74** | **0.74** | **0.76** | **0.74** |
| AdaBooster | **0.63** | **0.63** | **0.63** | **0.63** |
| GaussianNB | **0.52** | **0.40** | **0.61** | **0.51** |

**Area Under the Curve (AUC)**

Area under ROC curve is often used as a measure of quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1.

* .90-1 = excellent (A)
* .80-.90 = good (B)
* .70-.80 = fair (C)
* .60-.70 = poor (D)
* .50-.60 = fail (F)

**Accuracy(CA):** the proportion of the total number of predictions that were correct.

**Precision:** the proportion of positive cases that were correctly identified.

**Sensitivity or Recall**: the proportion of actual positive cases which are correctly identified.

**Model Choosen : Random Forest**

**Accuracy of Predicitons:**

1.Data Without Past History: 73.9%

2.Data With Past History : 93.1%