Originally, the dataset is consist of 466,566 samples, 13 numerical features and 34 non-numereical features with one being the response class. We first dropped the columns as\_of\_year, state\_name, state\_abbr, agency\_name, msamd\_name, county\_name, for redundancy, rate\_spread for too many NAs, which was ensued by removing the following columsn for irrelevancy to our objective: respondent\_id, application\_date\_indicator, sequence\_number, purchaser\_type\_name, preapproval\_name,edit\_status\_name. The feature rate\_spread was excluded as well for more than 95% of it was NA and hence no sensible way for imputation could be found. Next, since our analysis only concerns the primary market where borrowers and lenders are involved, we removed the 48,356 rows where the action taken is "Loan purchased by financial institution". Those 60, 358 unfinished load applications indicated by label “Application withdraw by applicant” and 18,176 labeled “File closed for incompleteness” were also excluded because we were only interested in the complete ones ready for review. We also took out those 52 loan that would go through pre-approval considerations, hence who were labeled as “Preapproval request approved but not accepted” and “Preapproval request denied by financial institution” to further focus on the regular loans. As a result, the 11,735 “Application approved but not accepted” and 263,712 “Loan originated” were combined as the approved class, whereas the 64,177 “Application denied by financial institution” were considered as the other denied class. Finally, we eliminate another 264 rows contains NAs to make the whole dataset NA-free without any imputation technique. As the last step, we would combine the information contained in applicant\_ethnicity\_name, applicant\_race\_name\_1, applicant\_race\_name\_2, applicant\_race\_name\_3, applicant\_race\_name\_4 and applicant\_race\_name\_5 together down to a new feature named as applicant\_race. In the process, if the applicant tended to self-claim more than 1 race group, they would be relabeled as a new class as “claimed mixed”. The identical procedure would be implemented on the co\_applicant related columns and denial\_reason columns to condense all of the relevant information into two new feature named as co\_applicant\_race and denials. By this point, we would end up with a dataset of 311, 774 completed loan applications with 251,758 being approved and 60, 016 denied. There are 8 numeric features: tract\_to\_msamd\_income, population, minority\_population, number\_of\_owner\_occupied\_units, number\_of\_1\_to\_4\_family\_units, loan\_amount\_000s, hud\_median\_famlity\_income, applicant\_income\_000, as well as another 12 non-numeric feature: property\_type\_name, owner\_occupancy\_name, loan\_type\_name, loan\_purpose\_name, lien\_status\_name, hoepa\_status\_name, co\_applicant\_sex\_name, sensus\_tract\_number, applicant\_sex\_name, agency\_abbr, applicant\_race, co\_applicant\_race.