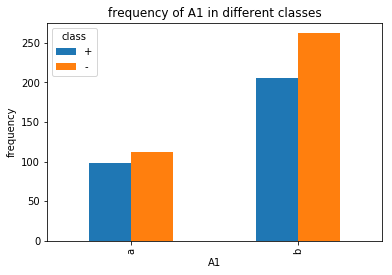
The given data consist of 44.49% approved credit cards records and 55.50% non-approved credit cards records. Data consists of total 678 records

In data, variables A1, A2, A4, A6, A7, A14 contain missing values. Since variables A1, A4, A6, A7 are categorical, hence missing values are replaced with most frequent value in variables and variables A2, A14 are numeric, hence replaced with median

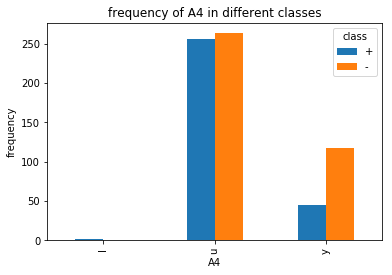


Variable ‘A1’ consists of two categories category ‘a’ and category ‘b’.

67.82% of records consists of category ‘b’ and 30.43 % of records consists of category ‘a’

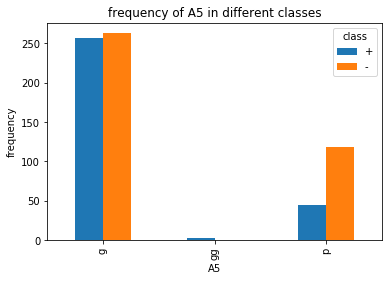
Total count of a: 210 and Total count of b: 468.

Out of 67.82% records of category ‘b’, 46.66 % records have been approved for credit cards and 53.33% of records have not been approved for credit cards. Out of 30.43% records of category ‘a’, 46.67% records have been approved for credit cards and 53.33% of records have not been approved for credit cards.



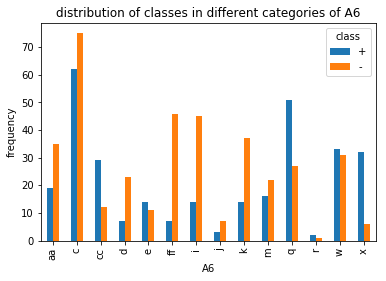
Variable ‘A4’ consists of two categories category ‘l’, category ‘u’ and category ‘y’.

The frequency of category if very less. The proportion of credit card approved is less than credit card not approved in category y. and the proportion of credit card approved is almost same as credit card not approved in category u and category l contains only 2 records.



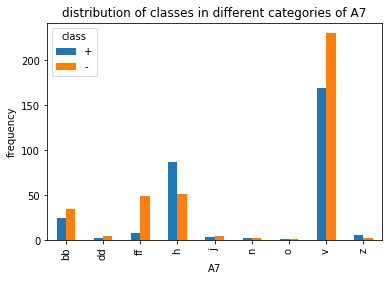
Variable ‘A5’ consists of two categories category ‘g’, category ‘gg’ and category ‘p’.

The frequency of category gg is very less. The proportion of credit card approved is less than credit card not approved in category p. and the proportion of credit card approved is almost same as credit card not approved in category g. Category ‘gg’ contains only 2 records.



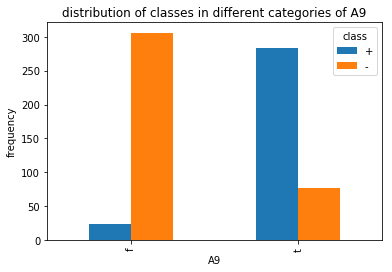
Variable ‘A6’ consists of different 14 categories category as aa, c, cc, d, e, ff, I, j, k, m, q, r, w, x.

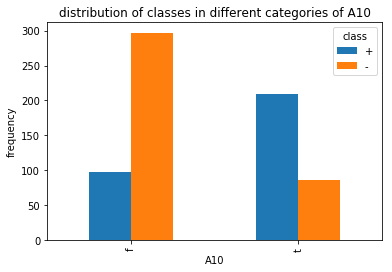
In categories, cc, e, q, w, x, r the proportion of credit card approved is more than credit card not approved and in categories aa, c, d, ff, i, j, k, m the proportion of credit card approved is less than credit card not approved.



Variable ‘A7’ consists of different 9 categories category bb, dd, ff, h, j, n, o, v, z.

In categories, h, z the proportion of credit card approved is more than credit card not approved and in categories bb, dd, ff, j, v the proportion of credit card approved is less than credit card not approved. Frequency of categories dd, j, n, o, z is very negligible.

Variable ‘A9’ consists of different 2 categories category f and t.

The proportion of credit card approved is less than credit card not approved in category f and the proportion of credit card approved is more than credit card not approved in category t.

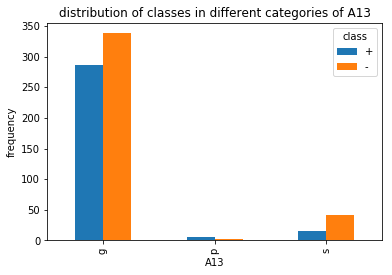
Variable ‘A10’ consists of different 2 categories category f and t.

The proportion of credit card approved is less than credit card not approved in category f and the proportion of credit card approved is more than credit card not approved in category t.



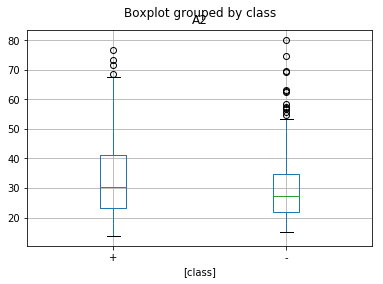
Variable ‘A12’ consists of different 2 categories category f and t.

The proportion of credit card approved is less than credit card not approved in both categories f and t.

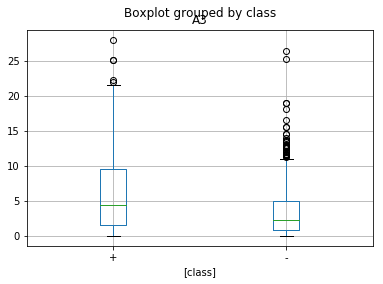


Variable ‘A13’ consists of different 3 categories category g, p and s.

The proportion of credit card approved is less than credit card not approved in categories g and s. Though frequency of category p is very, but proportion of credit card approved is more than credit card not approved. But as compared to category g, frequency of category p and s is very low.



Both distributions have roughly the same center (medians are 30.5 for credit card approved (+), and 27.33 for credit card not approved(-)). Also, A2 in Class + have roughly same variability as A2 in class – (range in += 63 , range in -=65.08).On other hand, if we look at the IQR, which measures the variability only among the middle 50% of the distribution, we see more spread in A2 for class + (IQR=Q3-Q1=18.16) as compared to A2 in class – (IQR=12.83). We see that we have outliers in both distributions. But there is less outlier in class + as compared to class -. Hence conclude that, A2 of class + are more consistent than A2 of class- which varies a lot. However, the middle 50% of the A2 distribution of class- is more homogeneous than the class +’s A2.

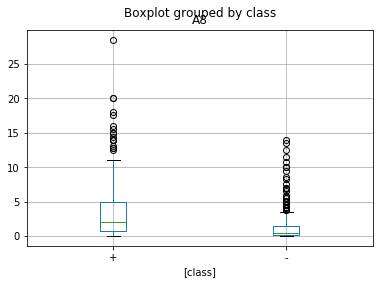


Median (+): 4.46 Median (-):2.21

IQR (+): 8.02 IQR (-):4.165

Range (+): 28 Range (-):26.335

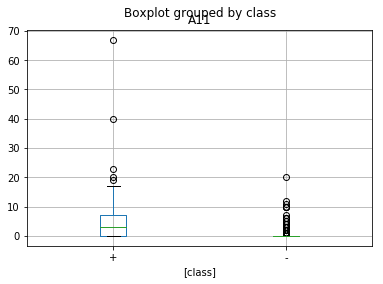
From here we can see that, distribution of class + is homogeneous that than distribution of class- for middle 50% of data. But overall, distribution of A3 is varies a lot in class- than that of class+ (more outlier in class- as compared to class+).



Median (+): 2 Median (-):0.415

IQR (+): 4.25 IQR (-):1.375

Range (+): 28.5 Range (-):13.875

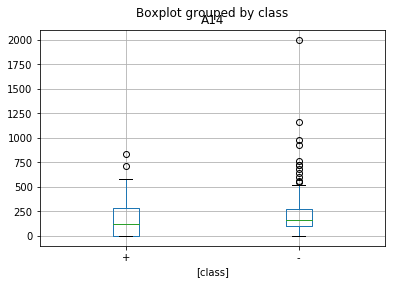
Here we can clearly see that, distribution of A8 for class – is more homogenous than distribution of A8 for class +.

Median (+): 3 Median (-):0

IQR (+): 7 IQR (-):0

Range (+): 67 Range (-):20

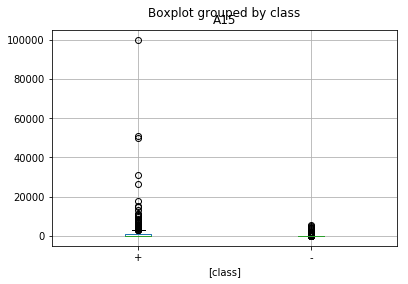
Here we can clearly see that, distribution of A11 for class – is consistent as compared to distribution of A11 for class+.



Median (+): 120 Median (-): 167.5

IQR (+): 280 IQR (-):172

Range (+): 840 Range (-):2000

Here we can see that, distribution of A14 for class + is more consistent as compared to distribution of A14 for class-

Median (+): 221 Median (-): 1

IQR (+): 1209 IQR (-):67

Range (+): 100000 Range (-):5552

Feature engineering:

1. Remove records with category ‘l’ in variable A4.
2. Remove records with category ‘gg’ in variable A5.
3. If categories are present in less than 5% then combine them together and categories them in category rare, since if data is split

Feature selection:

Chi square test is used to find association between independent categorical and dependent categorical variables. After applying chi-square test, came to know that variable ‘A1’ and ‘A12’ do not have association with dependent variable ‘class’. Since p value is not less than 0.05

Hence, not used for model building.

Model is built using logistic regression, decision tree and random forest. And Decision tree is given better result as compared to other 2 algorithms with below accuracy.

Train

Accuracy= 0.896049896049896

Precision= 0.8785046728971962

Recall= 0.8867924528301887

F1\_score= 0.8826291079812206

Test

Accuracy= 0.8840579710144928

Precision= 0.8709677419354839

Recall= 0.8709677419354839

F1\_score= 0.8709677419354839

Also, here we can see that model is not overfitted.