Determining whether a person is entitled to a loan

Miloš Stanković SW50-2016

1. Motivation

I wanted to create an application that will determine based on several attributes whether a person has the right to get a loan or not.

2. Research questions

The main question in this research is which attributes influence the decision of whether or not someone can get a loan and to what extent. Attributes that are in the data set are: Loan id, gender, whether the person is married or not, number of dependents, education, self employed or not, applicant's income, coapplicant's income, loan amount, where does the applicant live and his credit history.

There are 614 entries in the training data set and 367 entries in the testing data set.

3. Related work

While researching about this topic I found a few similar projects and the one thing that most of them had in common is that they used applicant's income and number of dependents as their main attributes of interest.

4. Methodology

I used Logistic regression and SVM while solving this problem and then compared the results.

5. Discussion

The first thing I did after loading the data set was to check the percentages for all attributes. For example, in the training set there are: 69% allowed loans and 31% denied.

Then I went through every attribute and made a graph so that I can see how does the data in that category affect whether on not the person is getting the loan.

The first thing I noticed is that people who had loan history are very likely to get another loan.

The next one was Marriage and I noticed that people who are married have higher chance to get a loan than people who are not married.

While looking at education, gender and self employment I didn’t find any useful connections.
While looking at number of dependents I noticed that if it either zero or two there is a higher chance that the person is getting a loan than if it is one or 3+.
The next one is the area where the person lives, and I noticed that if the person is living in the semiurban area than they have the highest chance of getting the loan.
While looking at applicant’s income I didn’t find any connection but the higher the coapplicant’s income is the higher the chance for loan is.

After finishing with that part I decided to split the data set into two, one with attributes that are filled with numerical values and one with attributes that are filled with categorical values. Then I filled the missing data in both (for numericals I decided to copy the value from the one before him, and for categorical I took the most popular values).
Then I combined the data sets again and trained them with both Logistic regression and SVM. Regression gave the accuracy of 79,96% while SVM gave the accuracy of 69,06% which leads to a conclusion that for this particular problem Regression is better than SVM.