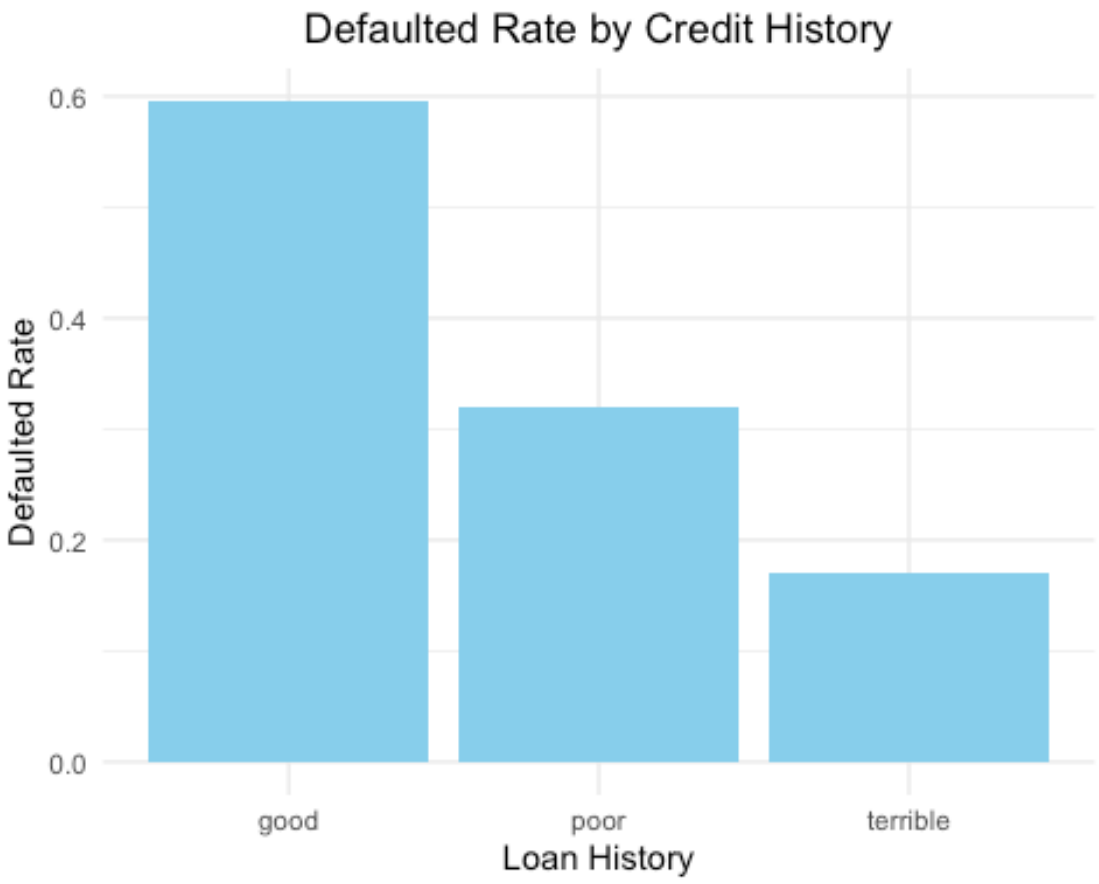
hw2_q2

2024-02-25

2. Classification and retrospective sampling

In problem 2, we are dealing with a data set from a German bank including information about loans. Our goal is to investigate how predict default probability is related to some other characteristics of defaulted loans, and try to make predictions based on information we have.



The bar plot above describes the defaulted rate categorized by different levels of credit history. There are three levels of history: “Good”, “Poor”, and “Terrible”. By observing the plot we can see that among three levels of credit history, the loans with “Good” history have the highest defaulted rate, while the loans with “Terrible” history have the lowest defaulted rate. This result is counter-intuitive, because it suggests that better credit history is related to higher loan defaulted rate.

Now we are going to build a prediction model with logistic regression and see if it is a good prediction of the defaulted rate.

Coefficients reported by logistic regression model:

```
##         (Intercept)            duration              amount
installment2
##               -0.42                0.03                0.00
0.08
##        installment3        installment4                 age
historypoor
##                0.42                0.60               -0.02
-1.11
##     historyterrible          purposeedu purposegoods/repair
purposenewcar
##               -1.88                0.72                0.10
0.85
##      purposeusedcar       foreigngerman
##               -0.81               -1.28

##    yhat
## y     0   1
##   0 674  26
##   1 256  44
```

According to the confusion matrix: error rate = (256+26)/1000=0.282, which indicates 72% accuracy. This is not a very high accuracy rate.

Although the coefficients to some extent indicate reasonable relationship between some characteristics, for example, the installment with defaulted probability, we can still see counter-intuitive factors, as well as an unsatisfactory accuracy rate of 72%. Combining the regression result and the bar plot, we can reasonably make a hypothesis that there's something in the data which prevents us from making successful predictions.

Calculate counts of samples falling into different categories:

```
## [1] "Number of 'good' credit history =  89"

## [1] "Number of 'poor' credit history =  618"

## [1] "Number of 'terrible' credit history =  293"
```

Here we can see a huge gap between counts. That is to say, oversampling of some certain categories in the data may potentially be the reason why counter-intuitive statistical results occur. The loans with "good" credit history are underrepresented in the data, and a large portion of them happen to be defaulted loans. This problem may have been caused by how the data was originally selected. Since the loans in the data set was manually selected based on whether the loans have similar situations as the defaulted loans, these loans in the data set cannot represent the real life distribution of borrowers.

Thus, this data set is an inappropriate one for building a predictive model for defaults. To classify borrowers into "low" and "high" defaulted probability categories, the bank needs a data set which more accurately represents the real distribution and situations of the potential borrowers. For example, randomized sampling can be a good way to achieve this.