LendUp Data Challenge

Jie (Jessica) Zhu (zhjie@umich.edu)
10/22/2015

1. An example of the output of your function from Question #1 for a single column

(a) Example of numerical feature: dti

Statistics:
dti
Mean: 17.059
Std : 7.597
NAN Count: 0 [0.0%]

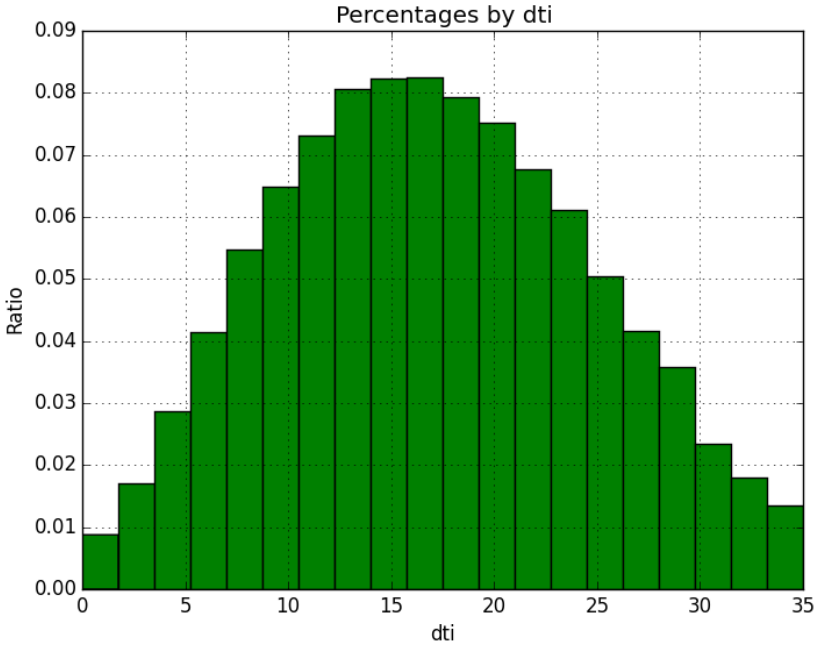

Figure 1. Distribution of dti in the dataset

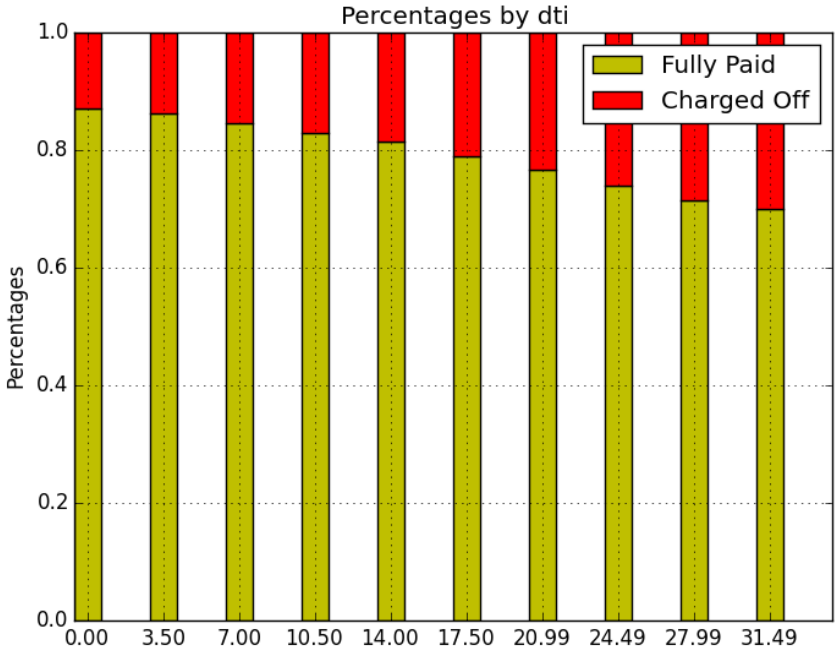

Figure 2. Ratio of dti by loan status between Fully Paid or Charged Off

(b) Example of categorical feature: grade

Statistics:
grade
A: 0.152
B: 0.333
C: 0.266
D: 0.148
E: 0.065
F: 0.030
G: 0.006
NAN Count: 0 [0.0%]

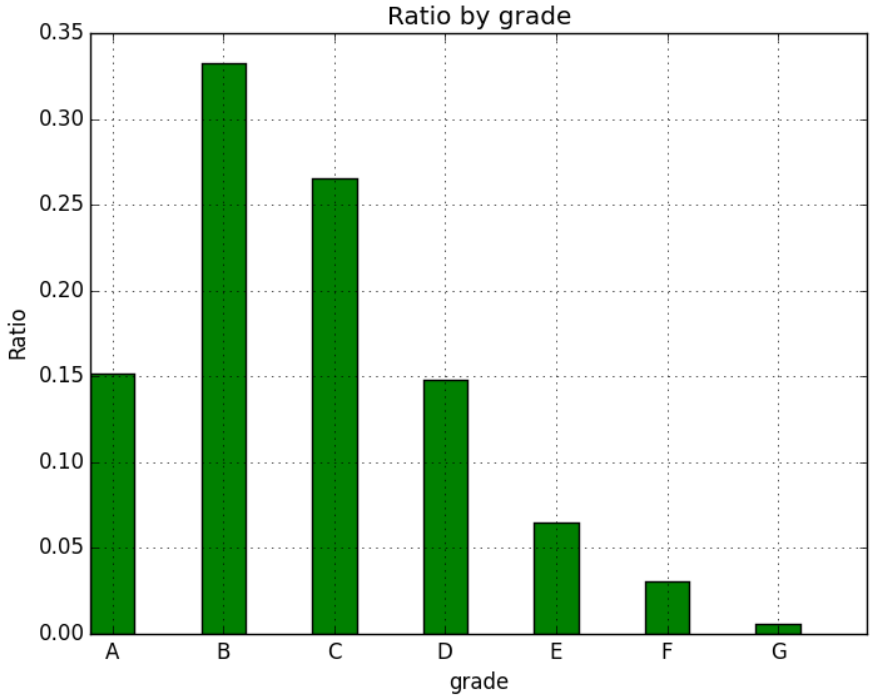

Figure 3. Distribution of grade in the dataset

Figure 4. Ratio of grade by loan status between Fully Paid or Charged Off

2. Your argument, along with any tables or graphs from Question #2

Preprocess:

- Separated the dataset in to a training set and a test set by loan status as:

  Training set: Fully Paid (positive) and Charged Off (negative)
  Test set: Current, In Grace Period, Late, Default

- Found that the model tended to cause bias the training set if using some accumulative quantities; Excluded these features in the model:

  'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'recoveries', 'collection_recovery_fee'.

- Discarded time features because they introduced time bias.

- Performed binary encoding for categorical features (such as grade, term, zip_code, et al.).

Model:

Xgboost: learning rate: 0.05; number of iterations: 100; depth = 9

Cross validation:

Training AUC score: 0.85
Test AUC score: 0.71

Feature Importance by Random Forest

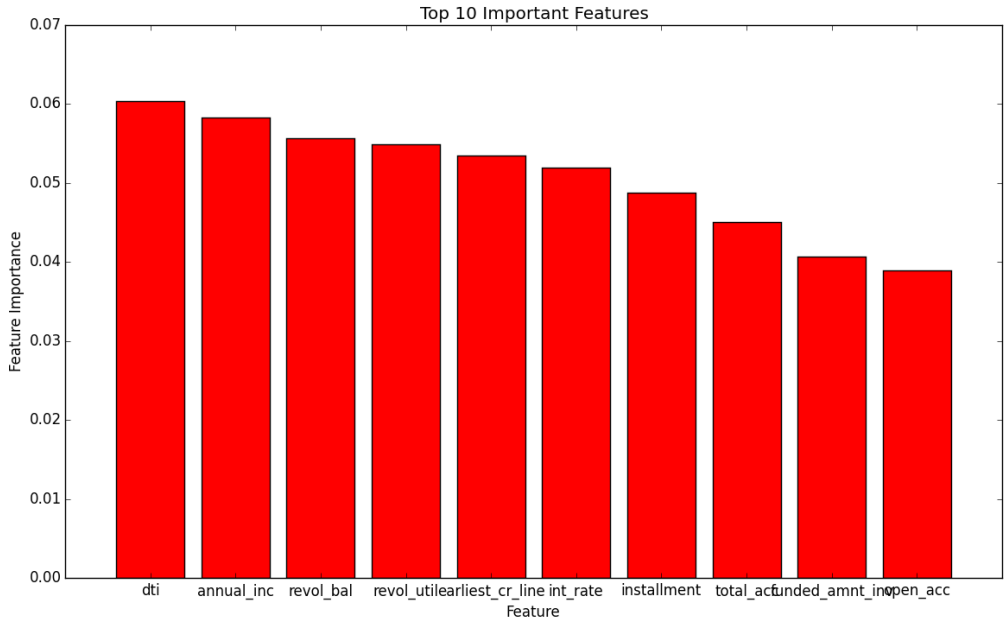

Figure 5. Top ten important features

Ratio of Fully Paid based on current status

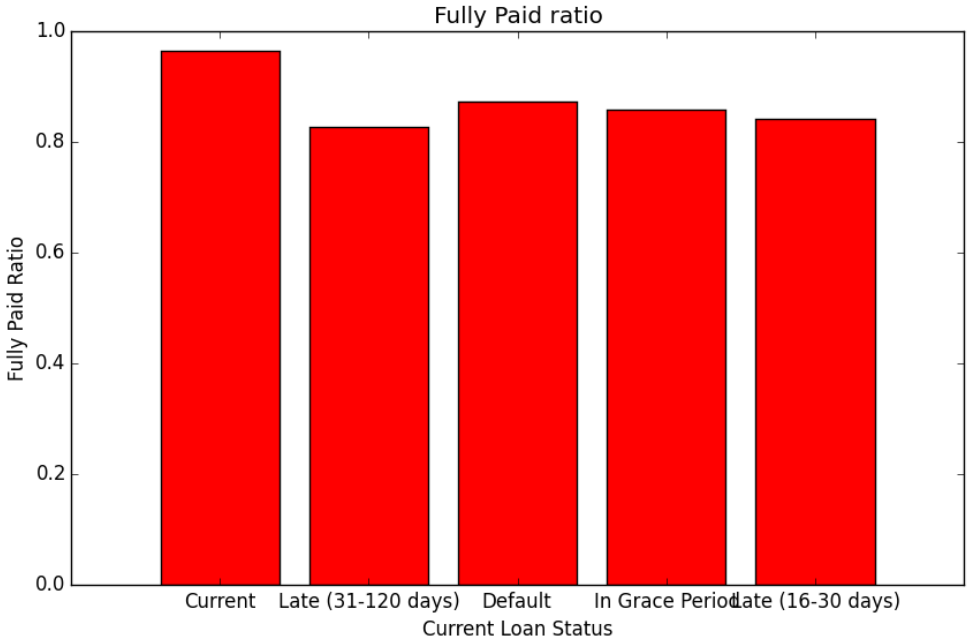

Figure 6. Fully Paid ratio by current loan status

Discussion

(a) Regardless of features that tend to cause bias, the most important feature is dti: a ratio calculated using the borrower's total monthly debt payments on the total debt obligations, divided by the borrower's self-reported monthly income.
(b) About 95% of the loans of the current borrowers are expected to be fully paid; about 80% of the loans of other loan status are expected to be fully paid.
(c) The predicted fully paid ratio for the test set (80%~95%) is higher than that in the training set (80%), which may be caused by not using those accumulative quantities. However, further investigation or more data is needed for address this problem.
(d) Prediction is save prediction in 'prediction.csv'