Credit Approval Analysis

Ryan Kuhn

May 1, 2015

Background

Objective:
To demonstrate the analytical techniques taught during the Special Topics in Audit Analytics course at Rutgers University

Data:
- Dataset- Credit Screening on Credit Card Applications
- Source- UCI Machine Learning Repository

Assumptions:
Field names and values changed to meaningless values. Made assumtions about what attributes the data represents.

Research Question - 1

Q: Is there a relationship between Age, Income, Credit Score, and Debt levels and the credit approval status? Can this relationship be used to predict if a person is granted credit? If yes, does the relationship indicate reasonable risk management strategies?

Research Question - 1

Q: Is there a relationship between Age, Income, Credit Score, and Debt levels and the credit approval status? Can this relationship be used to predict if a person is granted credit? If yes, does the relationship indicate reasonable risk management strategies?

A: Relationships exists between Prior default, Years employed, Credit score, and Income level. These variables are reasonable management strategies.

Research Question - 2

Q: Ethnicity is a protected status and the decision to approve or deny an application cannot be based on the applicant's ethnicity. Is there a statistically significant difference in how credit is granted between ethnicities that could indicate bias or discrimination?

H0: Ethnicity and approval are independent.
H1: Approval status is dependent on the ethnicity and credit card company has a compliance risk.

Research Question - 2

Q: Ethnicity is a protected status and the decision to approve or deny an application cannot be based on the applicant's ethnicity. Is there a statistically significant difference in how credit is granted between ethnicities that could indicate bias or discrimination?

H0: Ethnicity and approval are independent.
H1: Approval status is dependent on the ethnicity and credit card company has a compliance risk.

A: A chi-squared test did not give evidence that ethnicity and approval status are dependent. We cannot reject the null hypothesis.

Analytic Methods Used

Methods used:

- Linear regression
- Descriptive Statistics and Normalization
- Association Rules
- Logistic regression
- Classification and Regression Tree
- Ensembling

Linear Regression

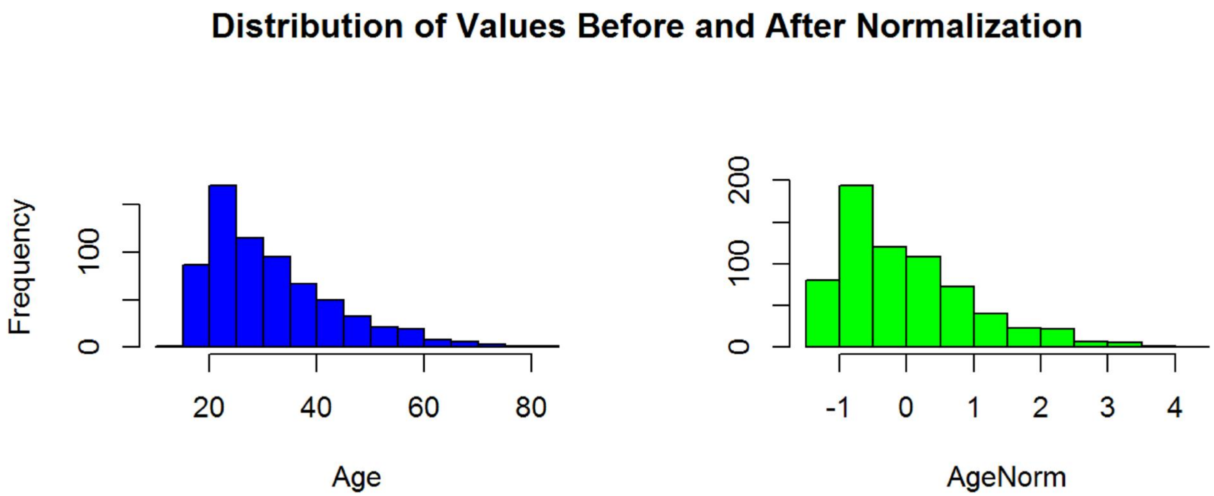
Used to fill in missing values in Age

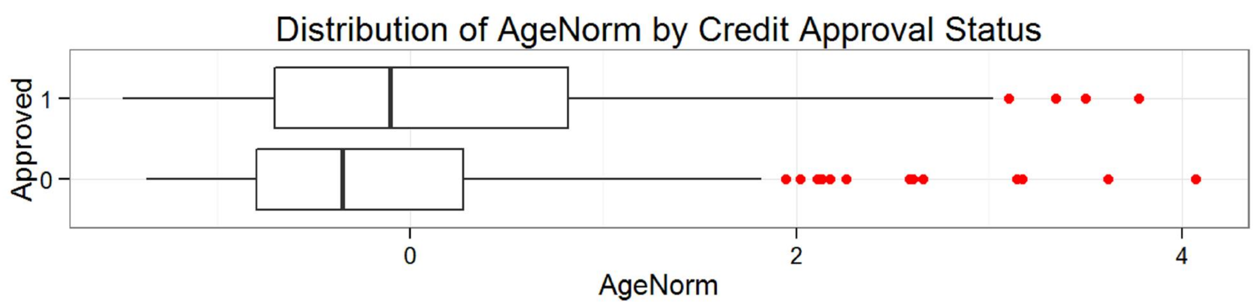
	Age	Debt	YearsEmployed	CreditScore	Income
Age	1.000	0.202	0.396	0.186	0.019
Debt	0.202	1.000	0.301	0.271	0.122
YearsEmployed	0.396	0.301	1.000	0.327	0.053
CreditScore	0.186	0.271	0.327	1.000	0.063
Income	0.019	0.122	0.053	0.063	1.000

```
  (Intercept) YearsEmployed
    28.446953      1.412399
```

Descriptive Stats and Normalization

Convert values to Z-Scores





Association Rules

1. Support: Support is how often the left hand side of the rule occurs in the dataset. In our example above, we would count how many times {u,g,c} occurs and divide by the total number of transactions.

2. Confidence: Confidence measures how often a rule is true. First, we find the subset of all transactions that contain {u,g,c}. Of this subset, we then count the number of transactions that match the right hand side of rule, or {1}. The confidence ratio is calculated by taking the number of times the rule is true and dividing it by the number of times the left hand side occurs.

```
  lhs                    rhs        support confidence     lift
1 {EducationLevel=c} => {Male=0} 0.1545319  0.7647059 1.099673
```

Baseline Model

- Simple mean of results.
- Establish benchmark to measure model accuracy
-

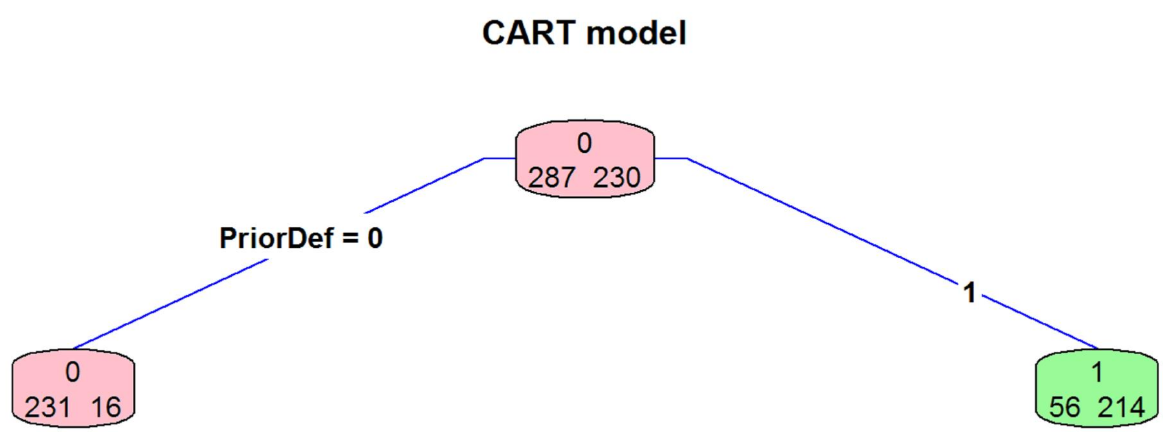
```
  Total Success Percent
1   517     230   44.49
```

Logistic Regression

```
Call:
glm(formula = Approved ~ AgeNorm + DebtLog + YearsEmployedLog +
    CreditScoreLog + IncomeLog, family = binomial, data = Train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4339  -0.7848  -0.4970   0.7164   2.1596

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.10596    0.11283  -0.939 0.347648
AgeNorm           0.01541    0.11674   0.132 0.894965
DebtLog           0.08430    0.11635   0.725 0.468744
YearsEmployedLog  0.70937    0.13021   5.448 5.10e-08 ***
CreditScoreLog    1.01687    0.13944   7.293 3.04e-13 ***
IncomeLog         0.46610    0.12089   3.855 0.000116 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 698.11  on 506  degrees of freedom
Residual deviance: 500.25  on 501  degrees of freedom
```

CART Model

```
n= 517

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 517 230 0 (0.55512573 0.44487427)
  2) PriorDefault=0 247  16 0 (0.93522267 0.06477733) *
  3) PriorDefault=1 270  56 1 (0.20740741 0.79259259) *
```



References

Data:
http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening

Analytic Report:
http://www.rpubs.com/kuhnrl30/CreditScreen