LendingClub Dataset Predictive Model

Aime\_Mukena

1/7/2020

Context & Content

This dataset comprises 9,578 entries(rows) and 14 columns. The dataset was provided by Kaggle and contains information on loan borrowers collected by LendingClub from the year 2007 to 2010. In this report, we are going to try to build the best predictive model possible to predict what type of borrower is more likely to pay back their loan. We will try to use as little statistical jargon as possible for easy understanding. However, a basic understanding of statistics is nonetheless required to fully grasp the content of this project.

You will find below a brief explanation of the column names used in the dataset.

1. credit.policy: a set of guidelines and criteria that decides on credit limits, credit terms, and how to go about delinquent accounts. The borrower either meets the criteria (represented by the integer 1 in our data) or does not (represented by 0 in our data).
2. purpose: purpose of the loan contracted (a.i: installment, debt consolidation)
3. int.rate: interest rate on the loan.
4. installment: monthly payment owed by the borrower.
5. log.annual.inc: natural logarithm of self-reported income.
6. dti: Debt-to-Income ratio
7. fico: FICO score of loan borrowers
8. days.with.cr.line: number of days with line of credit.
9. revol.bal: revolving balance or the portion of credit that remains unpaid after billing cycle.
10. revol.util: revolving utilization or debt-to-limit ratio, debt divided by credit limit.
11. inq.last.6mths: inquiries during last 6 months. Inquiries happen when a financial institution checks your credit to make a lending decision, this usually happens when you apply for credit. There are two types of inquiries: hard inquiries or ‘hard pulls’ when you apply for mortgage, credit card… and soft inquiries or ‘soft pulls’ for credit card offers or employment. For the purpose of this project, we will consider the data to represent ‘hard pulls’.
12. delinq.2yrs: number of loan delinquencies (30 days or more past due on a payment) reported during the past 2 years.
13. pub.rec: number of derogatory public records of loan borrowers.
14. not.fully.paid: this column shows whether a loan was fully paid (represented by the integer 1) or not fully paid (represented by 0)

Exploratory Data Visualization

We are starting our analysis with an exploratory data visualization of the dataset. This will allow us to have a sense of what’s happening.

01000200030004000all\_othercredit\_carddebt\_consolidationeducationalhome\_improvementmajor\_purchasesmall\_business

purpose

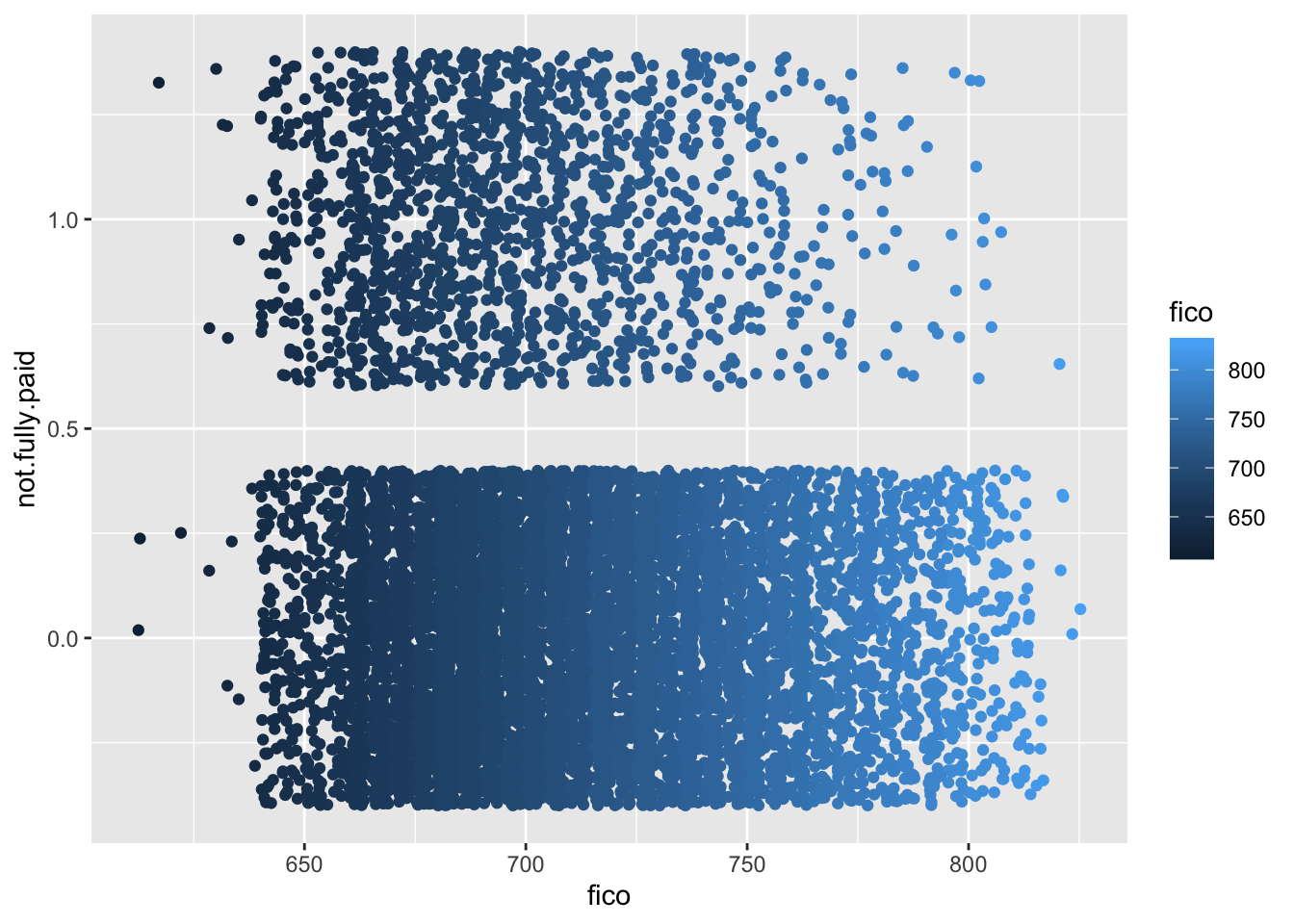
The first figure is a bar chart of purpose against payment status(paid or not paid). From the bar chart we can see that most people borrow money to consolidate a debt or pay off a credit card\*.

\*Unspecified payments classified as ‘all other’ are in fact the second largest portion of loans borrowed however we do not know in details what they represent.

A screenshot of a cell phone

Description automatically generated

The second figure is a plot of purpose against payment status(‘fully paid’ or ‘not fully paid’). From this plot, we can infer that ‘purpose’ is a good predictor as it quite unambiguously shows that purpose of the loan impacts payment status. However, this does not necessarily mean that whether a loan will be paid back or not depends on the purpose. It only confirms the existence of a relationship.



The third figure represents status of payments against FICO scores. From this plot we can infer that FICO scores can be good a predictor as well because individuals with a score inferior to 750 seem to make up the largest number of insolvents.

Statistical Analysis: Variable Distribution

In this section, we are going to look at how much variability there is in our data. We are going to plot some histograms in order to inspect our data for its distribution, outliers, skewness, etc.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA picture containing bird

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Statistical Analysis: Logistic Regression Model

Since we are dealing with classification (1 for ‘fully paid’ and 0 for ‘not fully paid’) and not a normal distribution pattern, a linear regression model would not be appropriate, hence the need to use a logistic regression model. We will try to interpret the output of the logistic regression below.

##

## Call:

## glm(formula = not.fully.paid ~ purpose + credit.policy + +fico +

## inq.last.6mths, family = "binomial", data = loan\_df)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -1.8950 -0.6084 -0.5067 -0.3953 2.5378

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 5.527593 0.617704 8.949 < 2e-16 \*\*\*

## purposecredit\_card -0.448835 0.106357 -4.220 2.44e-05 \*\*\*

## purposedebt\_consolidation -0.174372 0.073295 -2.379 0.0174 \*

## purposeeducational 0.137778 0.149967 0.919 0.3582

## purposehome\_improvement 0.096468 0.123546 0.781 0.4349

## purposemajor\_purchase -0.376618 0.165010 -2.282 0.0225 \*

## purposesmall\_business 0.732842 0.108920 6.728 1.72e-11 \*\*\*

## credit.policy -0.386173 0.080137 -4.819 1.44e-06 \*\*\*

## fico -0.009877 0.000895 -11.035 < 2e-16 \*\*\*

## inq.last.6mths 0.077393 0.013242 5.845 5.08e-09 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 8424.0 on 9577 degrees of freedom

## Residual deviance: 7958.9 on 9568 degrees of freedom

## AIC: 7978.9

##

## Number of Fisher Scoring iterations: 5

After trying multiple feature combinations - in other words, we ran a regression with different variables to find good predictors for our model, we now have the best feature combination consisting of purpose of loan + credit policy + FICO scores + hard pulls from the last 6 months with the status of payment as our dependent variable. All these features are statistically significant due to their p-value being inferior to α=.05. This means that we can use these features to predict loan payments.

Interpretation of Coefficients

1. Each one-unit change in purpose(for credit card) will decrease the log odds of payment status by -0.36, and its p-value indicates that it is significant in determining payment outcome.
2. Each one-unit change in fico will decrease the log odds of payment status by -0.0098, and its p-value indicates that it is significant in determining payment outcome.
3. Each one-unit change in inq.last.6mths will increase the log odds of payment status by 0.08, and its p-value indicates that it is significant in determining payment outcome.
4. Each one-unit change in credit.policy will decrease the log odds of payment status by -0.32, and its p-value indicates that it is significant in determining payment outcome.

Prediction

Our aim here is to predict the probability of a borrower to pay back their loan given his/her profile. Let’s consider a borrower with a relatively low FICO score of 580 who has been subjected to 1 hard pull in the last 6 months who wants to pay off a student loan. We will try to predict the chances that he/she will pay back their loan in its entirety.

## 1

## 0.4080246

We can predict that there’s a 40% chance that this particular borrower will pay back their loan in its totality.

Our second Predictive Model: Decision Tree

A picture containing object, clock

Description automatically generated

In the snapshot above, you can see that variable inq.last.6mths is the best predictor of loan payment. The nodes will split based on number of inquiries and FICO scores. If the borrower was subjected to 4 or more inquiries in the last 6 months, we will then look at their FICO score; if their FICO score is lower than 740, chances are they will not pay back their loan. If the borrower was NOT subjected to 4 or more inquiries, we will again look at their FICO score; less than 660 means that they will not pay back their loan while greater than 660 means that there are high chances that they will.

Comparing our Models

Lastly, we are going to compare the two predictive models.

A close up of a map

Description automatically generated

In the snapshot above, we are looking at a Cumulative Gains and Lift chart which allows us to assess the effectiveness of predicitive models. The blue line represents our logistic regression and the red line represents the decision tree. We use the Area Under the Curve(AUC) to measure the quality of the models. There is no overlapping, in fact, there is perfect seperation of the two lines and we can clearly see that the logistic regression model is the better model.