STAT GR5293: Financial Technology and Data-Driven Innovation - Problem Set #1

Xiaoli Sun (xs2338) - xs2338@columbia.edu

September 21, 2020

Problem 1: Exploratory Data Analysis

(a)

There are 10469 rows and 11 columns.
These functions in Python are used to get the answer.

```
count_row = data_36.shape[0]
count_col = data_36.shape[1]
```

(b)

Answer: as inTable 1
Python function used:

```
data_36.dtypes
```

(c)

Answer: as inTable 1

(d)

The distributions of Original Loan Balance, Pre-Loan DTI, and Original FICO are shown in Figures 1, 2, and 3 below.
For other answers, use the table 2 below.
Python function used:

```
sns.distplot(data_36['Original FICO'])
len(data_36['Original FICO'].unique())
```

Table 1: Question(b & c)

Field Name	(b) Data Type	(b) Numerical/Categorical?	(c) Units
Original Loan Balance	float	Numerical	dollar
Interest Rate	float	Numerical	ratio
Term	int	Numerical	month
Monthly Payment	float	Numerical	dollar
Annual Income	float	Numerical	dollar
Pre-Loan DTI	float	Numerical	The debt-to-income ratio
Original FICO	int	Numerical	Credit score (per point)
Number of Trade Lines Opened (Last 12 Months)	int	Numerical	counts per unit trade
Employment Length	float	Numerical	year
Housing Status	str object	Categorical	NA
Loan Status	str object	Categorical	NA

Table 2: Question(d)

Field Name	(d) Bounds?	(d) Continuous/Discrete?	(d) Number of Unique Va
Annual Income	No upper bound. Has lower bound.	Continuous	1317
Pre-Loan DTI	has both bounds	Continuous	3482
Original FICO	No upper bound. Has lower bound.	Discrete	38

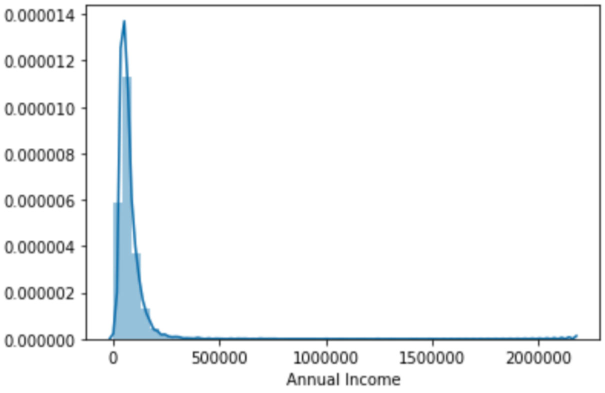


Figure 1: Histogram of Annual Income

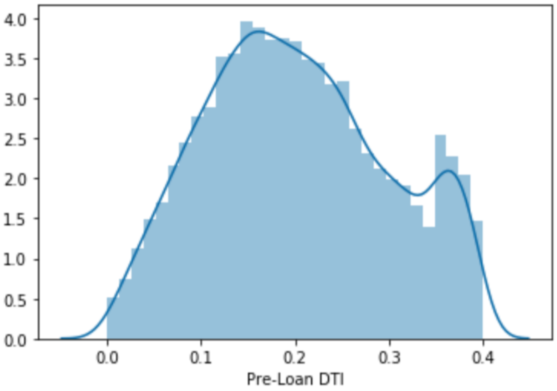


Figure 2: Histogram of Pre-Loan DTI

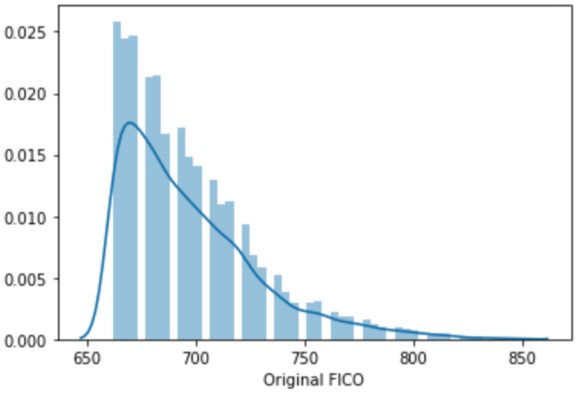


Figure 3: Histogram of Original FICO

Problem 2: Timeline Associated with Data

(a)

Each row represents a loan, and includes attributes of borrowers when they apply, decisions made by the lender or platform, and then what happens over the life of the loan. Describe the chronological order of how these fields are recorded in this dataset. That is,

(i) Which fields are recorded when a borrower applies for a loan with LendingClub?

Annual Income , Pre-Loan DTI , Original FICO , Employment Length , Housing Status .

(ii) Which fields are gathered when LendingClub inquires about a borrower when they contact credit bureaus and FICO?

Number of Trade Lines Opened (Last 12 Months) .

(iii) Which fields are set when LendingClub sets the terms of the loan?)

Original Loan Balance , Term , Interest Rate , Monthly Payment .

(iv) Which fields are determined some time after the borrower begins to pay back the loan (or not)?

Loan Status

(b)

Each of the attributes reflects the state of the borrower at the time of an application, but borrowers are dynamic! At least two of these fields change when a new loan is taken out. Please say which ones and describe, in mathematical terms, how they change.

answer:

new Number of Trade Lines Opened (Last 12 Months) = old Number of Trade Lines Opened (Last 12 Months) +1

$newPreLoanDTI = [oldPreLoanDTI * MonthlyIncome + MonthlyPayment(thisnewloan)]/MonthlyIncome$

where Monthly Income = Annual Income/12. (this is fixed value) new FICO = FICO-function(updated parameters)

(c)

ManyMLcoursesthatusecreditdatashowloanoutcomesexclusivelyfallingintotwoclasses: default or no default, and a simplification also used in many industry applications. In reality, loans involve payments by borrowers over time and so the outcome of a single loan does not fall neatly into two categories. However, for the purposes of analyzing trends, we will show one common way of mapping sequences of payments into two aforementioned classes.

(i) What outcomes are shown in the original dataset?

Python used:

```
data_36['Loan Status'].unique()
```

Outcome:

'Paid Off', 'Current', 'Sold - Debt Sale', 'Charged Off', '60 - 89 Days Delinquent', '30 - 59 Days Delinquent', '$\geq$ 90 Days Delinquent', 'Late (16 - 29 DPD)'

8 kinds of status.

(ii) Using the mapping shown in the notebook, what is the overall default prevalence?

Python used:

```
data_36_default['Default'].mean()
```

overall default prevalence = 0.03648868086732257. Prevalence of default is low. Most people make loan payment on time.

Problem 3: Understanding Original FICO, Pre-Loan, DTI, and Income

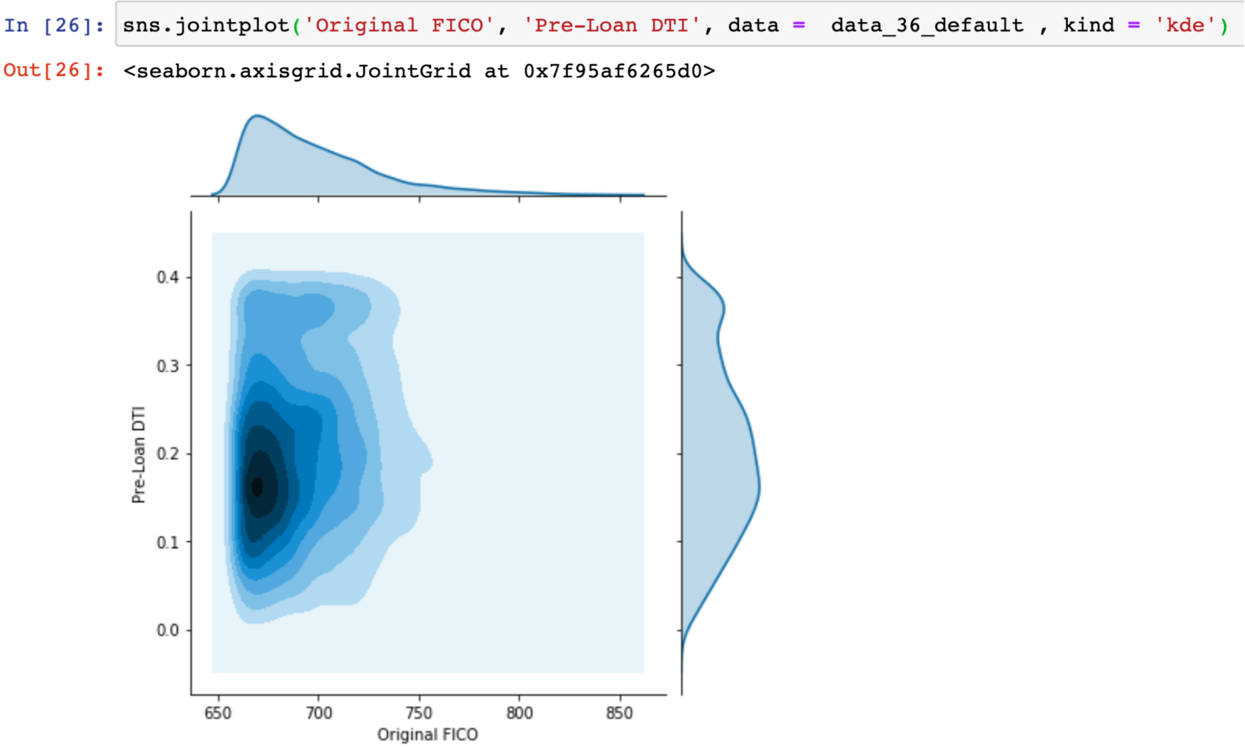
a

Use seaborn to create a joinplot (contour plot with histograms) of Original FICO scores ($x1$), Pre-loan DTI ($x2$), and prevalence (i.e. percentage) of default (y). Include this plot in your solutions write-up. Which variable–original FICO or pre-loan DTI–appears to be a better predictor of probability of default? Why? Which distribution is more symmetric?

Answer:

Python used:

```
sns.jointplot(’Original FICO’, ’Pre-Loan DTI’, data =  data_36_default , kind = ’kde’)
```



```
In [26]: sns.jointplot('Original FICO', 'Pre-Loan DTI', data =  data_36_default , kind = 'kde')
Out[26]: <seaborn.axisgrid.JointGrid at 0x7f95af6265d0>
```

Figure 4: Plotting FICO, DTI, and prevalance

pre-loan DTI appears to be a better predictor of probability of default. There are too many extreme points in the distribution of Original FICO. Original FICO has a long tail, which makes it a worse predictor.

pre-loan DTI has a more symmetric distribution.

(b)

(i)

How does the prevalence of default behave as a borrower's pre-loan DTI increases?
Decrease → Increase → Decrease

How does the prevalence of default behave as a borrower's original FICO score increases?
Overall decrease monotonicly.

(ii)

The population more evenly distributed between the bins for **pre-loan DTI** .
Higher proportions of population in each bin make me **more confident** about the associated prevalence of default. When the population bin grows taller, Original FICO becomes smaller, which implies that the borrower is not financially healthy. They are more likely to default.