Business Analytics' Final Assignment

"

Classification Project for Debt Non-payment Through Credit Score Prediction

"

21500268, Seongchan Park

Contents

Part 1

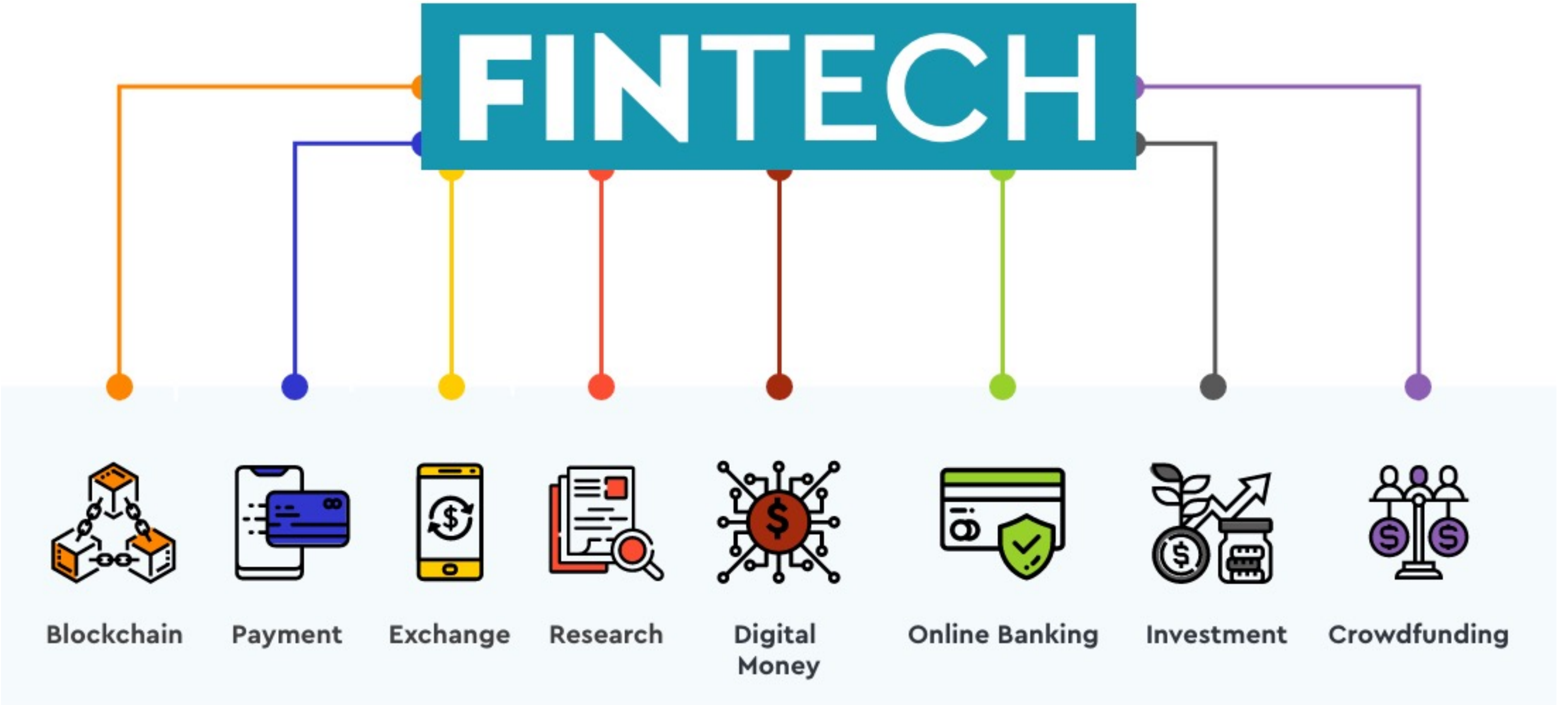
Background _ Self introduction

- I am majoring in ICT convergence and management.
- I am interested in solving many problems in our daily lives with technology based on insights through data. 'Finance' is one of the areas of my greatest interest.
- I am interning this semester at this Fintech-related company.

한국금융솔루션

1. **Loan comparison service**
2. Deposit account recommendation service
3. Asset allocation service using 'Mydata' of users
4. Robo-advisor service

Background _ Introduction of interest fields



Fintech is a combination of ‘Finance’ and ‘Technology’,
which is a financial service that utilizes advanced information technology,
and it is one of the most closely related parts of our lives.

Part 1

Background _ Introduction of interest fields

Classification Projects to Predict
Probability of Debt Non-Payment

(Defaulting)

Fintech is a combination of "Finance" and "Technology",
which is a financial service that utilizes advanced information technology,
and it is one of the most closely related parts of our lives.

Background _ Introduction of project

Importance of Default Risk (Searat Ali et al., 2018)

The debtor is unable to fulfill interest or principal repayment as set forth in the contract.

Reason

Firm's future cash flow is not sufficient to cover interest payments.

Result

Individual productivity may be lowered, mental stress, and suicide may occur.

Solution

It should maintain a better governance mechanism and reduce the asymmetry of information.

Source:
1. The Economic Terms Dictionary (Ministry of Strategy and Finance)
2. Searat Ali et al. (2018). Does corporate governance quality affect default risk? The role of growth opportunities and stock liquidity.

Background _ Introduction of project (Goal)

Risk can be reduced by distinguishing people with repayment ability

1. A credit scores are evaluated in consideration of **financial conditions** such as consumers' income or debt (data).
2. The **loan rate or the limit can be presented** through the objectively evaluated credit scores.
3. The financial state of the client is grouped, and data is utilized for the users of the similar group and the information of the **new customer** can be inferred.

Background _ Introduction of project (Motivation)

Interested in the financial problems closely related to our lives

1. I wanted to do a project that **deals with data in the financial sector** that I was interested in.
2. I wanted to do **Fintech-related projects** because I would continue to work in the field of Fintech after graduation.
3. I am interested in the subject of credit score prediction itself, so I can find out **how the personal data of customers affects credit score.**

Part 2

Data Analysis _ Introduction of data (Kaggle dataset)

Give Me Some Credit

Improve on the state of the art in credit scoring by predicting the probability that somebody will experience financial distress in the next two years.

k https://www.kaggle.com/c/GiveMeSomeCredit/data

Variable Name	Description	Type
DEFAULT (Target)	**Experienced 90 days past due delinquency**	**Factor - T/F**
Credit_limit_on_debt	Total balance on credit cards / Sum of credit limits	Integer - Percentage
Age	Age of borrower in years	Integer
Delay_30_59_days	Number of times borrower has been 30~59 days past due	Integer
Debt_ratio	Monthly debt payments / Monthly gross income	Integer - Percentage
Monthly_income	Monthly income	Real
Num_open_credit_loans	Number of open loans and lines of credit	Integer
Delay_90_days	Number of times borrower has been 90 days or more past due	Integer
Num_open_mortage_loans	Number of mortgage and real estate loans	Integer
Delay_60_89_days	Number of times borrower has been 60~89 days past due	Integer
Num_dependents	Number of dependents in family excluding themselves	Integer

Source:
Kaggle - Give Me Some Credit

Understanding data & Checking basic statistics

```
'data.frame':	150000 obs. of  11 variables:
 $ DEFAULT              : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
 $ Age                  : int  45 40 38 30 49 74 57 39 27 57 ...
 $ Delay_30_59_days     : int  2 0 1 0 1 0 0 0 0 0 ...
 $ Delay_60_89_days     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Delay_90_days        : int  0 0 1 0 0 0 0 0 0 0 ...
 $ Num_open_credit_loans : int  13 4 2 5 7 3 8 8 2 9 ...
 $ Num_open_mortage_loans: int  6 0 0 0 1 1 3 0 0 4 ...
 $ Credit_limit_on_debt : num  0.766 0.957 0.658 0.234 0.907 ...
 $ Debt_ratio           : num  0.803 0.1219 0.0851 0.036 0.0249 ...
 $ Monthly_income       : int  9120 2600 3042 3300 63588 3500 NA 3500 NA 23684 ...
 $ Num_dependents       : int  2 1 0 0 0 1 0 0 NA 2 ...
```

'str' output in R

Type identification of variables

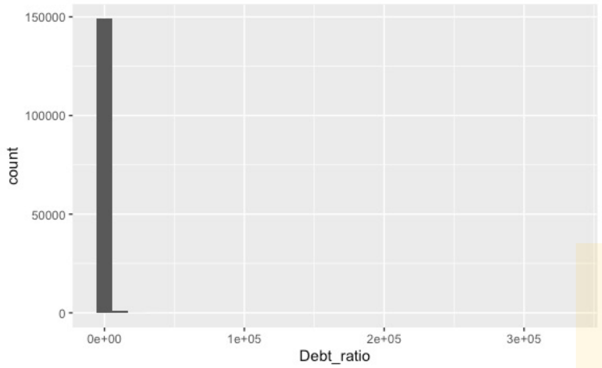
Variable	Min	1st_Qu	Median	Mean	3rd_Qu	Max
Age	21	41	52	52.29556	63	109
Delay_30_59_days	0	0	0	0.4210295	0	98
Delay_60_89_days	0	0	0	0.2403883	0	98
Delay_90_days	0	0	0	0.2659751	0	98
Num_open_credit_loans	0	5	8	8.452776	11	58
Num_open_mortage_loans	0	0	1	1.018233	2	54
Credit_limit_on_debt	0	0.02986692	0.1541758	6.048472	0.5590438	50708
Debt_ratio	0	0.1750736	0.3665032	353.0074	0.868257	329664
Monthly_income	0	3400	5400	6670.227	8249	3008750
Num_dependents	0	0	0	0.7572138	1	20

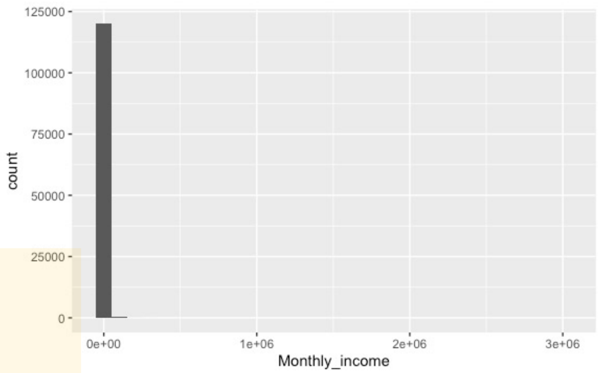
'summary' output in R

Statistics by each variables

Data Analysis _ Data preprocessing and EDA

Checking the normalization of data - Skewness

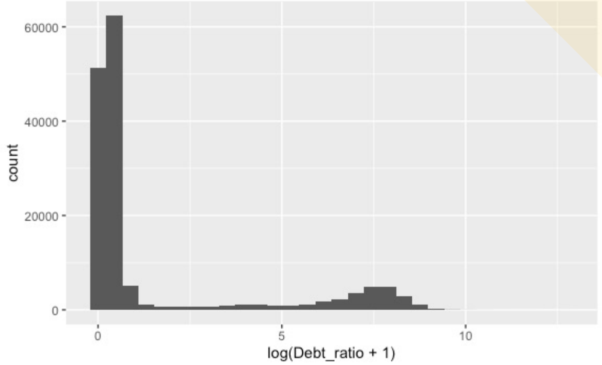


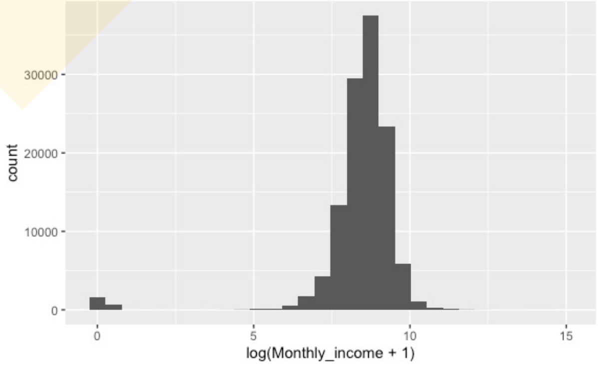


Direction and degree of distribution

```
               DEFAULT                    Age      Delay_30_59_days
             3.4687734              0.1892388            22.5965873
      Delay_60_89_days          Delay_90_days Num_open_credit_loans
            23.3311983             23.0868064             1.2152794
Num_open_mortage_loans    Credit_limit_on_debt            Debt_ratio
             3.4824420             97.6292964            95.1555976
        Monthly_income         Num_dependents
           119.9032853              1.6260549
```

Log normalization

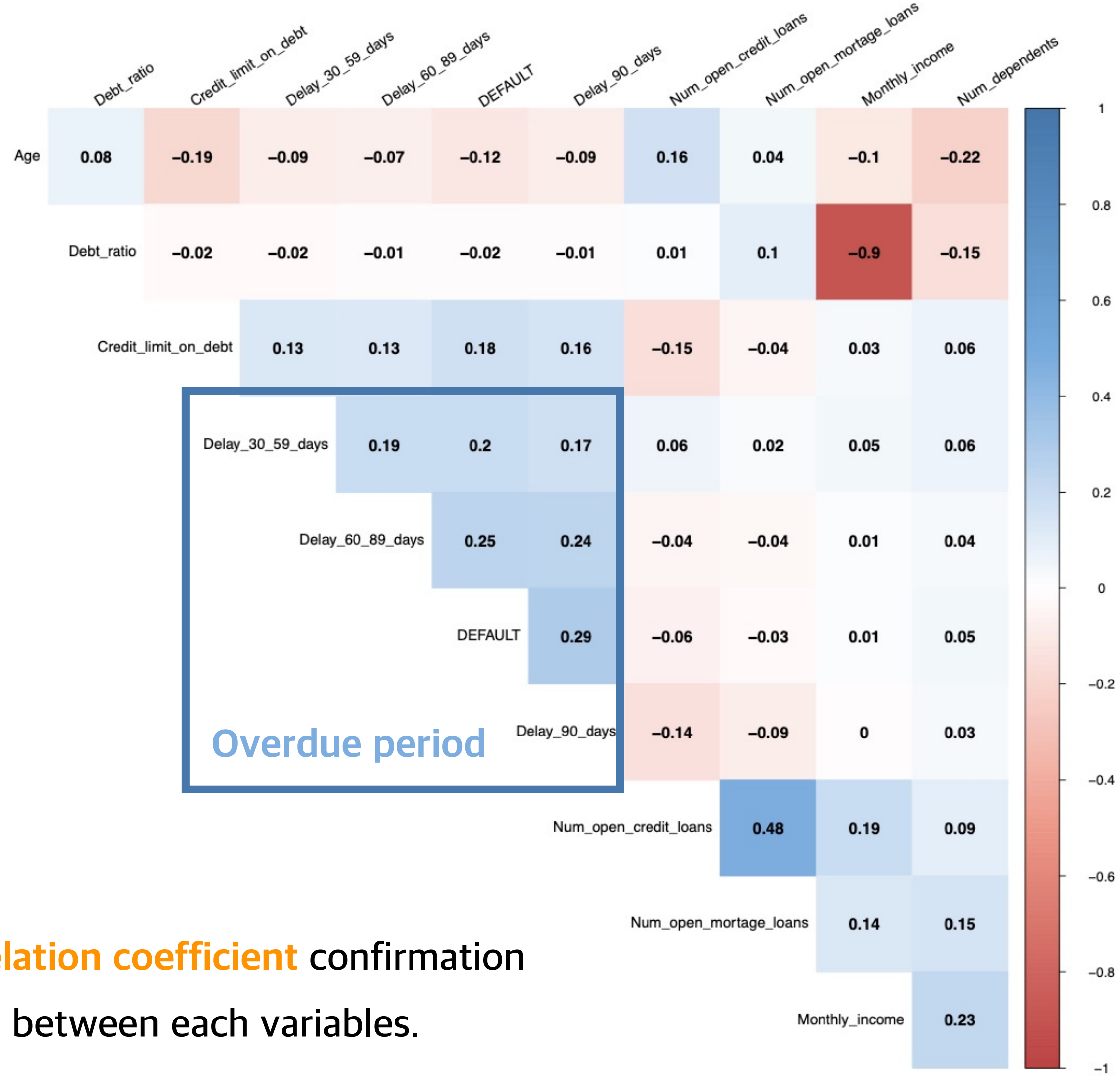





```
               DEFAULT                    Age      Delay_30_59_days
             3.4687734              0.1892388             2.0949949
      Delay_60_89_days          Delay_90_days Num_open_credit_loans
             4.3837731              4.3176369            -0.7330688
Num_open_mortage_loans    Credit_limit_on_debt            Debt_ratio
             0.2388435             11.7046881             1.7489676
        Monthly_income         Num_dependents
            -1.2917707              1.6260549
```

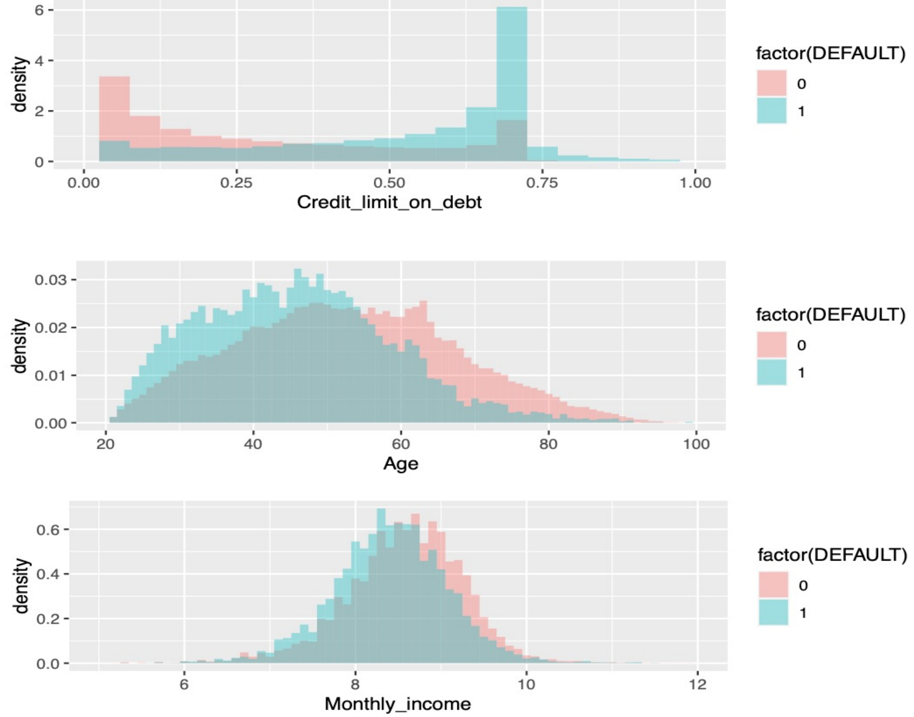
Part 2

Data Analysis _ Data preprocessing and EDA



Correlation coefficient confirmation

between each variables.



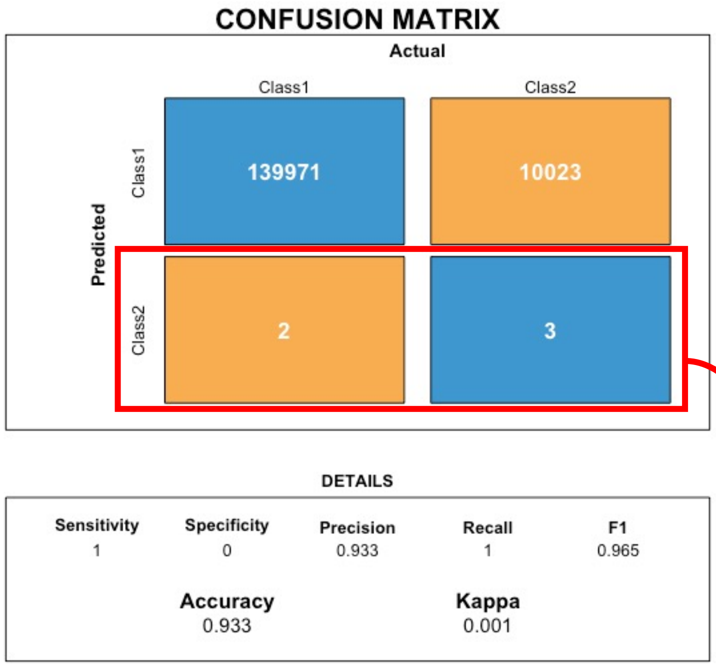
Check the distribution of each variables

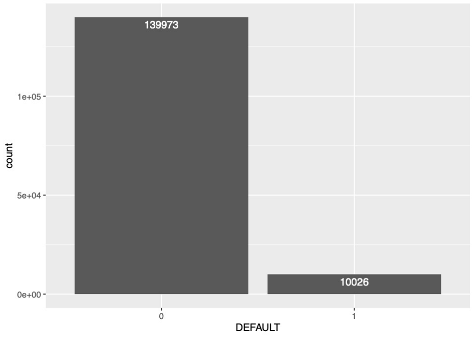
according to the target variable.

Part 2

Data Analysis _ Modeling and Performance (Logistic Regression)

```
Call:
glm(formula = DEFAULT ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9174  -0.3201  -0.2600  -0.2109   3.1079

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.0174008  0.0806769 -25.006  < 2e-16 ***
Age                    -0.0241950  0.0008899 -27.190  < 2e-16 ***
Delay_30_59_days        0.7997217  0.0186375  42.909  < 2e-16 ***
Delay_60_89_days        1.0231302  0.0231632  44.171  < 2e-16 ***
Delay_90_days           1.3369928  0.0230601  57.979  < 2e-16 ***
Num_open_credit_loans  -0.0799226  0.0222789  -3.587 0.000334 ***
Num_open_mortage_loans  0.1237271  0.0279371   4.429 9.48e-06 ***
Credit_limit_on_debt    0.5531399  0.0194673  28.414  < 2e-16 ***
Debt_ratio             -0.0334170  0.0123602  -2.704 0.006859 **
Monthly_income         -0.0214078  0.0094288  -2.270 0.023179 *
Num_dependents          0.0542008  0.0097915   5.536 3.10e-08 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 73616  on 149998  degrees of freedom
Residual deviance: 59931  on 149988  degrees of freedom
AIC: 59953

Number of Fisher Scoring iterations: 6
```





The model forecasts almost all debt fulfillment (0); **default (1) is not produced.**

=> The problem of Imbalanced data

Data Analysis _ Future plan (for final report)

(1) Imbalanced data

- Using method of over-sampling, under-sampling to solve data imbalance problem.

(2) Modifying input variables as a result of prior research

- Study on the influence of credit card and loan factors, the influence of age and monthly income, and credit rating.
- Generate and supplement new variables that can reflect this research.

(3) Confirming similar customer groups and proposing information using clustering

- Clustering generation according to customer's financial status (K-means, KNN)
- When a new customer comes in, recommend a similar cluster and present information.

Insights _ Impact and contribution

“ My data service, which started in December in Korea ”

Recommending products or services through own data
by checking **distributed data in various sector** at one.

매일경제 | A14면 TOP | 2021.11.25. | 네이버뉴스

"금융실적없는 1200만명 잡아라" 빅테크 전쟁

빅테크, 新금융 고객으로 공략 네이버, 선구매 후지불 서비스 토스, 자체 신용평가 만들기도 1200만명... 카드 실적이 없어도 신용도를 측정할 수 있게 되면서다. 네이...

Catch 12 million without financial performance:
Big Tech War

- Use other customers’ financial **databases to create clusters** of type.
- Customers who do not have enough data based on clusters can **estimate financial status** by using only a small amount of data.
- It may not be **discriminated against by the service** regardless of the amount of data.

Insights _ Impact and contribution

Micro

Increased understanding of credit status

Personalized service available to you

Minimize personal mental, physical damage

WITH

Macro

Potential risk reductions for lenders

Proper interest rates to be offered

Equal service activation available

Thank you :)