Final Project Report

IST687 - Group B2

9/26/2018

## Table of Contents

### INTRODUCTION

1. Project Background and Description
2. Project Scope and Context of this Analysis

### BUSINESS QUESTIONS

1. What are the Business Questions?

### DATA ACQUISITION, CLEANING, TRANSFORMATION, MUNGING

1. Describe your data acquisition process
2. What data did you select, all, subset, why
3. What was your initial quality assessment
4. What fields/variables did you finally decide on, why
5. Provide a data dictionary
6. Provide data descriptive statistics, rows, str
7. Did you have to do any cleansing, describe
8. Interesting findings

### DESCRIPTIVE STATISTICS

1. Provide demographic statistics – Location
2. Any early observations, nuggets of interest, interpretation, interesting findings
3. Graphs, charts, tables, visuals, text

### USE OF MODELING TECHNIQUES

1. Linear modeling
2. Support vector
3. Provide key statistics of interest and interpretation for each model

### OVERALL INTERPRETATION OF RESULTS/ACTIONABLE INSIGHTS

### REFERENCES

### APPENDIX – RStudio CODE

# FINAL PROJECT REPORT

## IST687 - GROUP B2

### INTRODUCTION

1. Project Background and Description  
   This project is an exercise in taking a dataset with a variety of financial metrics to generate questions regarding best business practices and finding ideal customers that would be approved to receive a financial loan.
2. Project Scope and Context of this Analysis  
   The scope of this project encompasses data gathered from a Financial Technology (FinTech) Company on customer demographics, current debts, income, applied loan amount, applied duration of loan, loan date, and more.  
   The data also shows which customers had their loans approved and which didn’t; allowing us to look for patterns as to the ideal customer. We paid attention to each customers Credit Score, as well as his financial statements, noting that customers with a higher Credit Score are more likely to be approved for a loan, while customers with large amounts of debt and a low income are not likely to be approved.  
   The data has been encrypted so to protect customer identity.

### BUSINESS QUESTIONS

1. What are the Business Questions?  
   The main question we are trying to answer is what characteristics must a customer have in order to be approved - or rejected - for a loan. This question aims to build a better model that will automatically approve a customer for a loan based on the data at hand. Another question that we have are which customers are more likely to inquire about taking out a loan. Answering this question will better help me understand were marketing efforts should be focused on in order to try and gain more customers that satisfy the demographic data.  
   Finally, a third question we would like to answer is how many loans can we expect to disburse over the next three months and how much will the month to month profit be.

### DATA ACQUISITION, CLEANSING, TRANSFORMATION, AND MUNGING

1. Describe your data acquisition process.  
   The data was acquired with permission from Martin’s workplace. It provides a data from customers who applied for a personal loan from September 2017 to January 2018. The data was anonymized in order to protect customer private information.  
   We loaded the data using the read\_excel() function from the read\_excel() package.

dat <- data.frame(read\_excel('IST687 Data.xlsx'))  
head(dat)

## person\_id credit\_score loan\_application\_date  
## 1 116 638 2017-09-26  
## 2 278 608 2017-10-23  
## 3 315 563 2017-10-26  
## 4 418 613 2017-11-07  
## 5 34 631 2017-09-07  
## 6 213 769 2017-10-12  
## requested\_loan\_term..in.months. requested\_loan\_amount..in.dollars.  
## 1 9 2000  
## 2 6 1200  
## 3 6 1700  
## 4 12 2000  
## 5 12 7000  
## 6 6 4200  
## preapproved\_loan\_amount..in.dollars. preapproved\_loan\_term..in.months.  
## 1 4900 12  
## 2 1200 6  
## 3 1700 6  
## 4 1300 6  
## 5 7000 12  
## 6 2200 6  
## person\_birthdate person\_marital\_status person\_gender  
## 1 1966-10-19 Married MALE  
## 2 1987-02-02 Married MALE  
## 3 1953-01-20 Married MALE  
## 4 1982-08-11 Married FEMALE  
## 5 1958-08-01 Married MALE  
## 6 1988-04-08 Married MALE  
## person\_degree\_type\_desc person\_employment\_type\_desc industry  
## 1 Technical Employed Sales  
## 2 High School Diploma Employed Sales  
## 3 High School Diploma Employed Sales  
## 4 High School Diploma Employed Sales  
## 5 Undergraduate Employed Sales  
## 6 Undergraduate Employed Sales  
## estimated\_net\_income..in.dollars. actual\_net\_income..in.dollars  
## 1 1800 1779.692  
## 2 2500 2660.630  
## 3 2100 2083.760  
## 4 1300 1407.000  
## 5 13200 7999.842  
## 6 4000 3990.333  
## monthly\_debt\_capacity..in.dollars. decision\_status  
## 1 523.19 Approved  
## 2 647.87 Approved  
## 3 348.04 Approved  
## 4 273.29 Approved  
## 5 715.00 Approved  
## 6 420.38 Approved  
## loan\_disbursement\_date first\_repayment\_date loan\_amount..in.dollars  
## 1 2017-09-27 2017-11-02 4900  
## 2 2017-10-24 2017-11-14 1200  
## 3 2017-10-27 2017-12-07 1700  
## 4 2017-11-08 2017-12-07 1300  
## 5 2017-09-08 2017-10-07 7000  
## 6 2017-10-13 2017-11-23 2200  
## approved\_loan\_rate approved\_nominal\_rate approved\_interest\_amount  
## 1 44.96 38.10 1273.52  
## 2 78.47 66.50 289.11  
## 3 61.74 52.32 318.91  
## 4 78.47 66.50 313.20  
## 5 39.18 33.20 1572.86  
## 6 43.57 36.90 287.88

1. What data did you select, all, subset, why?  
   We selected a subset of the data that encompasses customers who applied for a loan - regardless of whether they were approved for the loan or disbursed the loan - from September 2017 to January 2018. These months were selected as, by this point, the company had been active for six months and we wanted to check what kinds of customers were being attracted to the company or what kinds of loans - length and amount - they were asking for.
2. What was your initial quality assessment?  
   The first quality assessment was to check that there was no missing data within each column. NAs where replaced accordingly based on the column information. For example, for loan\_disbursement\_date, for those customers who weren’t approved or didn’t accept the load, NAs where replaced with 0 - or 01/01/1900 in Excel - to identify these customers. Fortunately, this was not the case for most of the columns like applied\_loan\_rate, loan\_disbursement\_amount which all had no missing data whatsoever.
3. What fields/variables did you finally decide on, why  
   We selected 23 variables which include a mixture of demographic data and loan data - allowing us to identify different customers, what possible background information they might have, and what type of loan they were seeking. Aside from person\_id, our demographic data included credit\_score, person\_birthdate (from which we calculated person\_age), person\_marital\_status, person\_gender, person\_degree\_type\_desc, person\_employment\_type\_desc, and industry.  
   The remaining columns were harbored information about the loan. For columns focused on the date - i.e loan\_application\_date, loan\_disbursement\_date, and first\_repayment\_date - the format was yyyy-mm-dd. Columns regarding the loan amount and other monetory data - i.e requested\_loan\_amount, preapproved\_loan\_amount, estimated\_net\_income, actual\_net\_income, monthly\_debt\_capacity, loan\_amount - the data is in US dollars. Data regarding the time it will take to pay off the loan - i.e requested\_loan\_term, preapproved\_loan\_term - is in months. Rate data is on a percentage base.  
   Though we feel that all this information adds insight on whether a person is approved for a loan, we will not be using all the available information to build our final model.

colnames(dat)  
names <- c("person\_id", "credit\_score", "loan\_application\_date", "requested\_loan\_term", "requested\_loan\_amount", "preapproved\_loan\_amount", "preapproved\_loan\_term", "person\_birthdate", "person\_marital\_status", "person\_gender", "person\_degree\_type\_desc", "person\_employment\_type\_desc", "industry", "estimated\_net\_income", "actual\_net\_income", "monthly\_debt\_capacity", "decision\_status", "loan\_disbursement\_date", "first\_repayment\_date", "loan\_amount", "approved\_loan\_rate", "approved\_nominal\_rate", "approved\_interest\_amount")  
colnames(dat) <- names

1. Provide a data dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Column Name | New Names | Definition |
| 1 | person\_id | Customer ID | Customer identification number. |
| 2 | credit\_score | Credit Score | Customer risk credit score. |
| 3 | loan\_application\_date | Loan Application Date | Date customer applied for a loan. |
| 4 | requested\_loan\_term (in months) | Requested Loan Term | Number of months customer has requested to pay off the loan. |
| 5 | requested\_loan\_amount (in dollars) | Requested Loan Amount | Amount customer wants to borrow. |
| 6 | preapproved\_loan\_amount (in dollars) | Preapproved Loan Amount | Amount customer was preapproved for. |
| 7 | preapproved\_loan\_term (in months) | Preapproved Loan Term | Number of months customer was preapproved for to pay off the loan. |
| 8 | person\_birthdate | Customer Birthdate | Customer date of birth. |
| 9 | person\_marital\_status | Customer Marital Status | Customer marital status. |
| 10 | person\_gender | Customer Gender | Customer gender. |
| 11 | person\_degree\_type\_desc | Customer Education Degree | Description of customer education degree. |
| 12 | person\_employment\_type\_desc | Customer Employment | Whether customer is dependently or independently employed. |
| 13 | industry | Customer Industry | Industry in which customer works. |
| 14 | estimated\_net\_income(in dollars) | Estimated Net Income | Estimated customer income after tax. |
| 15 | actual\_net\_income(in dollars) | Actual Net Income | Actual customer income after tax. |
| 16 | monthly\_debt\_capacity (in dollars) | Monthly Debt Income | Amount customer is able to pay on a monthly basis. |
| 17 | decision\_status | Loan Decision Status | Status of loan decision. |
| 18 | loan\_disbursement\_date | Loan Disbursement Date | Date when loan was taken out. |
| 19 | first\_repayment\_date | First Repayment Date | Date when customer should make first payment. |
| 20 | loan\_amount(in dollars) | Approved Loan Amount | Accepted loan amount by customer. |
| 21 | approved\_loan\_rate | Approved Loan Rate | Accepted loan rate by customer. |
| 22 | approved\_nominal\_rate | Approved Nominal Rate | Accepted loan nominal rate by customer. |
| 23 | approved\_interest\_amount | Approved Interest Rate | Accepted loan interest rate by customer. |

1. Provide data descriptive statistics, rows, str  
   The complete dataset has 23 variables and 852 observations. The dataset does not have any NA’s, meaning that there will be no additional data dropped.  
   The following shows the dimensions and structure of the dataset.

dim(dat)

## [1] 852 23

str(dat)

## 'data.frame': 852 obs. of 23 variables:  
## $ person\_id : num 116 278 315 418 34 213 441 712 131 226 ...  
## $ credit\_score : num 638 608 563 613 631 769 616 518 579 640 ...  
## $ loan\_application\_date : POSIXct, format: "2017-09-26" "2017-10-23" ...  
## $ requested\_loan\_term..in.months. : num 9 6 6 12 12 6 12 6 10 12 ...  
## $ requested\_loan\_amount..in.dollars. : num 2000 1200 1700 2000 7000 4200 5000 1500 2000 1800 ...  
## $ preapproved\_loan\_amount..in.dollars.: num 4900 1200 1700 1300 7000 2200 5000 1500 1300 1800 ...  
## $ preapproved\_loan\_term..in.months. : num 12 6 6 6 12 6 12 6 6 12 ...  
## $ person\_birthdate : POSIXct, format: "1966-10-19" "1987-02-02" ...  
## $ person\_marital\_status : chr "Married" "Married" "Married" "Married" ...  
## $ person\_gender : chr "MALE" "MALE" "MALE" "FEMALE" ...  
## $ person\_degree\_type\_desc : chr "Technical" "High School Diploma" "High School Diploma" "High School Diploma" ...  
## $ person\_employment\_type\_desc : chr "Employed" "Employed" "Employed" "Employed" ...  
## $ industry : chr "Sales" "Sales" "Sales" "Sales" ...  
## $ estimated\_net\_income..in.dollars. : num 1800 2500 2100 1300 13200 4000 10000 3000 3000 1800 ...  
## $ actual\_net\_income..in.dollars : num 1780 2661 2084 1407 8000 ...  
## $ monthly\_debt\_capacity..in.dollars. : num 523 648 348 273 715 ...  
## $ decision\_status : chr "Approved" "Approved" "Approved" "Approved" ...  
## $ loan\_disbursement\_date : POSIXct, format: "2017-09-27" "2017-10-24" ...  
## $ first\_repayment\_date : POSIXct, format: "2017-11-02" "2017-11-14" ...  
## $ loan\_amount..in.dollars : num 4900 1200 1700 1300 7000 2200 5000 1500 1300 1800 ...  
## $ approved\_loan\_rate : num 45 78.5 61.7 78.5 39.2 ...  
## $ approved\_nominal\_rate : num 38.1 66.5 52.3 66.5 33.2 ...  
## $ approved\_interest\_amount : num 1274 289 319 313 1573 ...

1. Did you have to do any cleansing, describe  
   No cleansing was needed as the dataset was already cleaned when collected from Martin’s company’s SQL repository. There were no errors in the dataset and no missing values.  
   However, we did need to add to columns to represent customer age, and whether the customer accepted the loan or not.
2. Interesting findings

### DESCRIPTIVE STATISTICS

1. Provide demographic statistics – Location
2. Any early observations, nuggets of interest, interpretation, interesting findings
3. Graphs, charts, tables, visuals, text

### USE OF MODELING TECHNIQUES

1. Linear modeling
2. Support vector
3. Provide key statistics of interest and interpretation for each model

### APPENDIX - RStudio Code

#Load the required packages  
packages <- c('pacman', 'readxl', 'dplyr', 'ggplot2', 'tidyr')  
install.packages(packages)  
require(pacman)  
p\_load(readxl, dplyr, ggplot2, tidyr)  
  
#Load the dataset into R   
dat <- read\_excel('IST687 Data.xlsx')  
  
#Check the structure of the data, head, and see that there are no NAs.  
str(dat)  
head(dat)  
summary(dat)  
  
#Column names need to be changed to better understand the data.   
colnames(dat)  
names <- c("person\_id", "credit\_score", "loan\_application\_date", "requested\_loan\_term", "requested\_loan\_amount", "preapproved\_loan\_amount", "preapproved\_loan\_term", "person\_birthdate", "person\_marital\_status", "person\_gender", "person\_degree\_type\_desc", "person\_employment\_type\_desc", "industry", "estimated\_net\_income", "actual\_net\_income", "monthly\_debt\_capacity", "decision\_status", "loan\_disbursement\_date", "first\_repayment\_date", "loan\_amount", "approved\_loan\_rate", "approved\_nominal\_rate", "approved\_interest\_amount")  
colnames(dat) <- names  
  
#Create month variable to better bin the data for month-to-month analysis.  
#We'll also calculate the age and bin them in groups of ten years.  
dat <- dat %>% mutate(age = as.numeric((Sys.Date() - as.Date(person\_birthdate))/365),  
 age\_bins = as.character(cut(age, breaks = c(21, 31, 41, 51, 61, 71, 81, 91),  
 labels = c('21-30', '31-40', '41-50', '51-60', '61-70', '71-80', '81-90'))),  
 loan\_application\_code\_month = ifelse(month(dat$loan\_application\_date) < 10, paste0(year(dat$loan\_application\_date), 0, month(dat$loan\_application\_date)), paste0(year(dat$loan\_application\_date), month(dat$loan\_application\_date))), loan\_disbursement\_code\_month = ifelse(month(dat$loan\_disbursement\_date) < 10, paste0(year(dat$loan\_disbursement\_date), 0, month(dat$loan\_disbursement\_date)), paste0(year(dat$loan\_disbursement\_date), month(dat$loan\_disbursement\_date))))  
  
#Demographic distributions  
#Applications per month by gender  
ggplot(dat, aes(loan\_application\_code\_month, fill = as.factor(person\_gender))) +  
 geom\_bar(position = 'stack') +  
 labs(x = 'YearMonth', y = 'Number of Applications', fill = 'Gender') +  
 ggtitle('Applications per Month', subtitle = 'Aug 2017 - Jan 2018') +  
 theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5))  
  
#Application per month by gender - percentage  
ggplot(dat, aes(loan\_application\_code\_month, fill = as.factor(person\_gender))) +  
 geom\_bar(position = 'fill') +  
 labs(x = 'YearMonth', y = 'Percentage of Applications', fill = 'Gender') +  
 ggtitle('Applications per Month', subtitle = 'Aug 2017 - Jan 2018') +  
 theme(plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5))  
  
#Applicant age distributions by gender  
ggplot(dat, aes(age, fill = as.factor(person\_gender))) +  
 geom\_density(alpha = 0.3) +  
 facet\_wrap(~ person\_gender, ncol = 1) +  
 labs(x = 'Customer Age', y = 'Density', fill = 'Gender') +  
 ggtitle('Age Distribution by Gender') +  
 theme(legend.position = 'bottom', plot.title = element\_text(hjust = 0.5), plot.subtitle = element\_text(hjust = 0.5))  
# Count by degree  
ggplot(dat, aes(reorder(person\_degree\_type\_desc, table(person\_degree\_type\_desc)[person\_degree\_type\_desc]))) +  
 geom\_bar() +  
 coord\_flip() +  
 labs(y = 'Customer Count', x = 'Degree')  
  
# Count by Line of work  
ggplot(dat, aes(reorder(industry, table(industry)[industry]))) +  
 geom\_bar() +  
 coord\_flip() +  
 labs(x = 'Industry', y = 'Count')  
  
# Count by marital status  
ggplot(dat, aes(reorder(person\_marital\_status, table(person\_marital\_status)[person\_marital\_status]))) +  
 geom\_bar() +  
 coord\_flip() +  
 labs(x= 'Marital Status', y = 'Count')  
  
# Credit score distribution  
ggplot(dat, aes(credit\_score)) + geom\_histogram()