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 <- read\_excel('IST687 Data.xlsx')

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.
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
4. Provide a data dictionary
5. Provide data descriptive statistics, rows, str
6. Did you have to do any cleansing, describe
7. 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()