Big Data for Finance

Final Project

CENSUS INCOME

Income range forecast

Caio Mescouto Terra de Souza
Loan NGUYEN-Thi-Anh
Vy NGUYEN-Phuc-Bao

March 8, 2021

Abstract

The main purpose of this report is to present three prediction models - supervised learning - applied to forecast income range through USA Census data from 1994/1995. The data was splited into training (2/3) and test (1/3). It was carried out a Logistic Regression model, a Decision Tree model and a Naïve Bayes and compared accuracy and the AUC. Our conclusion is that Logistic Regression was better in predict the income range.

Keywords: Machine Learning, Supervised Learning, Logistic Regression, Decision Tree, Naive Bayes, Big Data

Contents

Chapter 1

Introduction and Motivation

Introduction and Motivation

Researching Census income in a mainstream topic is carried out by a lot of authorities, national institutions as well as non-government organizations due to its wide and interactive applications. Census income is the foundation information providing detailed statistics that are essential for industries and communities. Trade associations, businesses and chambers of commerce rely on this data for economic development, business decision and strategic planning. These real-life applications urge us to carry out this interesting research with Census income data-set as well as extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau and implement BigData for Finance knowledge to provide, visualize and statistics information in order to solve the on-hand questions that we may have to handle in the professional career in the upcoming time.

In our work, we are making an attempt to establish a rule whereby we can classify a new individual (on the basis of observed attributes) into one of the existing classes and carry out the prediction based on observed attributes.

Specifically, we would like to have a summary of our analysis and exploration of Census income data to come up with meaningful, important and interesting attributes. After having sufficient knowledge about the attributes, we would perform a predictive task of classification to see whether the annual income of a new individual exceeds $50K. We also would like to describe the probability of an attribute of a new investigated resident, based on prior knowledge of conditions (features) related to that attribute. (For example, The probability of a person is a Native-born in the US given that his income is lower than $50K\year, his age is 40, he is a professor, etc…).

By practicing with this topic with such a huge data set, we hope that we can successfully apply and absorb the precious knowledge provided in the Big Data set course and truly understand how to skillfully apply such valuable knowledge to real-life topics.

Chapter 2

Data description

To serve our research purpose, we decided to employ Census-Income (KDD) Data set containing weighted census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau.

This big data-set contains 41 demographic and employment-related variables such as Age, Class of worker, Wage per hour, Sex, Citizenship, etc… There are 299285 observations while 199523 residents' information recorded in the data file and 99762 in the test file. One instance per line with comma-delimited fields.

The data set is quite tidy for studying purposes, we just need to do some codes to add the titles of attributes(the title of columns in the data frame) as it was absent in the original files and replace the sign "?" by NA, while the rest is well- organized and ready for analyzing progress.

Chapter 3

Descriptive Statistics

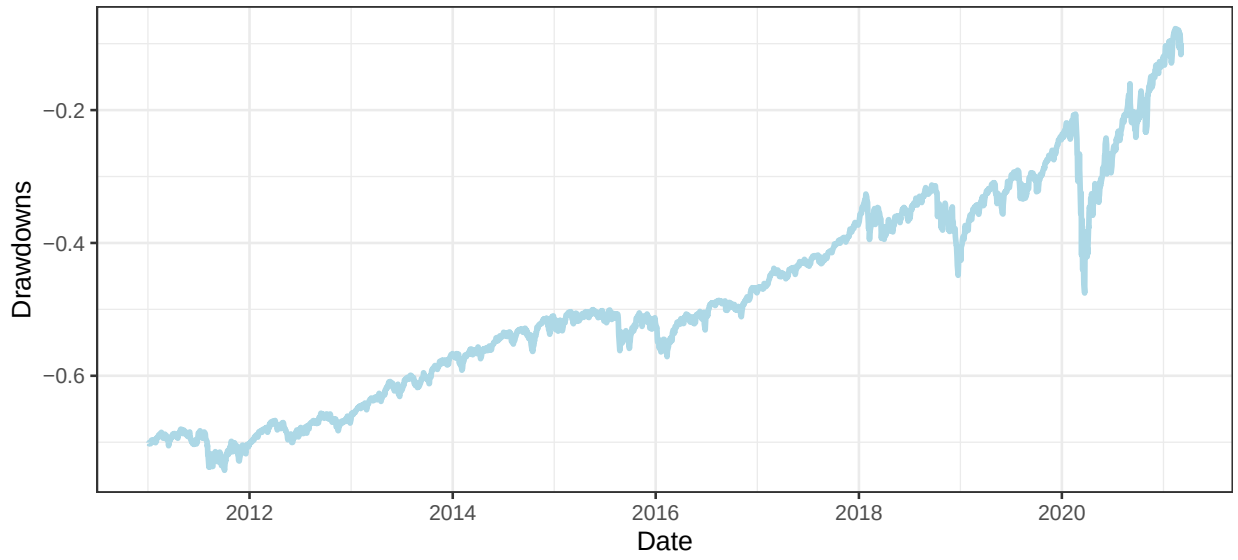

Figure 3.1: (#fig:fig_sp_500_drawdown)The S&P 500 suffered the Bigest drawdown for at least 10 years

Chapter 4

Methodology

cite (EMC 2015)

4.1 Logistic Regression

4.2 Decision Tree

4.3 Naïve Bayes

4.4 Confusion Matrix

4.5 ROC and AUC

Chapter 5

Results

Chapter 6

Conclusion

References

EMC, Education Services. 2015. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. 10475 Crosspoint Boulevard Indianapolis, IN 46256: John Wiley & Sons, Inc.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(data.table)
library(tidyverse)
library(R.utils)
library(rpart)
library(rpart.plot)
library(e1071)
library(lattice)
library(plyr)
library(caret)
library(ROCR)
library(tidymodels)
library(tidyquant)
library(dplyr)
library(tidyr)
library(ggplot2)
library(corrr)
library(gridExtra)
library(lme4)
library(car)
library(emmeans)

t_i = "2019-04-30"
t_f = "2019-12-30"

### Setting Event window.

t_event_i1 = "2020-02-18"
t_event_f1 = "2020-03-03"
t_event_1 = "2020-02-25"

### Gets data.

tickers = c("ZM", "^IXIC", "^VIX", "^GSPC")

price_data = tq_get(tickers) %>%
```

```
  group_by(symbol) %>%
  tq_mutate(select = adjusted,
            mutate_fun = periodReturn,
            period = "daily",
            col_rename = "daily_return") %>%
  mutate(wealth_index = cumprod(1*(1+daily_return)),
         previous_peaks = cummax(wealth_index),
         drawdowns = (wealth_index - previous_peaks)/previous_peaks)%>%
  select("date", "symbol", "adjusted", "daily_return", "wealth_index",
         "previous_peaks", "drawdowns")


yield_10 = tq_get(c("^TNX")) %>%
  mutate(adjusted = na.locf(adjusted, fromLast = FALSE)) %>%
  mutate(daily_yield = (1+adjusted/100)^(1/365)-1) %>%
  select("date", "symbol", "adjusted", "daily_yield")
# Plot: S&P 500 Drawdowns

ggplot(filter(price_data, symbol=="^GSPC"), aes(x=date, y=drawdowns))+
  geom_line(colour="Lightblue", size=1) +
  labs(x="Date", y="Drawdowns")+
  theme_bw(base_size = 10)
```