Contents

YSU FEM DSB Final Project Feedback - 1002/U91 |

Julieta, Hasmik

Coding

1. While running the code I faced with the errors of the absence of datasets.
   Error in mutate(Gcdaily, Gold = Close - lag(Close, default = Close[1]))
   : object 'Gcdaily' not found
   Solution is to call your files in codes:

```
Cldaily <- read.delim("Cldaily.txt", header = T,sep = ",")
Gcdaily <- read.delim("Gcdaily.txt", header = T,sep = ",")
eurusdDaily <- read.delim("eurusdDaily.txt", header = T,sep = ",")
```

2. You need to save the models in RDA format and then to load them on Shiny, or use the source function not to run the R script separately:

```
source("PROJECT R CODE _ Julieta, Hasmik.R")
load("PROJECT R CODE _ Julieta, Hasmik.rda")
```

3. Useless step, you can just compare the current values with past values:

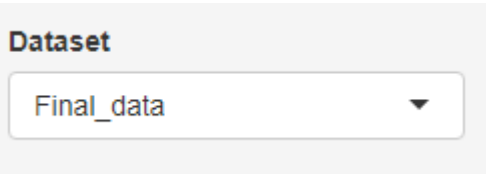
```
Gcdaily1 <- mutate(Gcdaily, Gold = Close - lag(Close, default = Close[1]))
Cldaily <- mutate(Cldaily, Oil = Close - lag(Close, default = Close[1]))


Gcdaily1 <- Gcdaily %>%
  mutate(Gold_dummy = ifelse(Gold > 0, 1, 0),
         Gold_dummy1 = ifelse(Gold < 0, -1, 0),
         Gold_dummy2 = Gold_dummy + Gold_dummy1)
```

4. There is no need to use 2 ifelse functions:

```
Gcdaily1 <- Gcdaily %>%
  mutate(Gold_dummy = ifelse(Gold > 0, 1, 0),
    Gold_dummy1 = ifelse(Gold < 0, -1, 0),
    Gold_dummy2 = Gold_dummy + Gold_dummy1)
```

5. Why do we need the drop down menu with one choice?



6. A lot of visually inappropriate typos e.g.:

Oil_dummy2	USD_dummy2	Gold_dummy2	AAPL_dummy2	FB_dummy2	AMZN_dummy2

Histogram of Final_data[[data1()]]

Conceptual

1. The histogram is for numeric data not categorical. The bar plot is more appropriate visualization tool for categorical data.
2. Independent variable can remain the same, and not to be changed to categorical. You lose the information.
3. Absence of models and interpretations in the report.
4. … nominal (equivalently categorical)… is not correct.

Questions

1. What does default mean?

```
lag(Close, default = Close[1])
```

2. Why sequential growth rate and not with the base?
3. Why did you choose the close price?
4. How the Amazon, Apple, Google stock prices are calculated?
5. How your ROC curve will be changed if you set 0 to 1, and vice versa?
6. How the specificity, sensitivity will be changed if you raise the threshold values.
7. How the logistic regression is estimated?
8. Interpret the coefficients.
9. Why OLS is not appropriate? (prediction, Bin, Hetero)
10. What is the range of p/(1-p)?
11. Show multinomial logit calculations for k=3
12. What is the non-information rate?

Summary

Criteria	%	Points	Julieta	Hasmik
Curiosity	20%	0.8	0.78	0.78
Skepticism	35%	1.4	0.8	0.8
Organization	25%	1	0.78	0.78
Shiny	20%	0.8	0.8	0.8
Total	100%	4	3.1	3.1

YSU FEM DSB Final Project Feedback - 1002/U91 |

Anna, Anush

Coding

1. I could not find anything new in your code.

2. It is not correct to use createDataPartition with numeric variable?

```
set.seed(1000)
training.samples <- df$residual.sugar %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data  <- df[training.samples, ]
test.data <- df[-training.samples, ]
```

3. Discordances in code output and interpretation, e.g:

```
## Residual standard error: 1.231 on 3908 degrees of freedom
## Multiple R-squared:  0.9403, Adjusted R-squared:  0.9401
## F-statistic:  5594 on 11 and 3908 DF,  p-value: < 2.2e-16
```

We see that adjusted R-squared of our model is 0.91. So if we didn't know that there is a correlation among the variables, we would think that we have a pretty good model. But that is not true.

4. The absence of sent data and Rmd file (I could not run your code).

Conceptual

1. The wrong usage of R^2 (you need to look at adjusted R^2 for multiple regression).
2. A lot of typos, copy-past formulas, difference in fonts, scientific notations which make your work less attractive.
3. The existence of incorrect statements e.g.:

   …The model will have a low accuracy if it is overfitting…
   …Multicollinearity is often described as the statistical phenomenon wherein there exists a perfect or exact relationship between predictor variables...

   - if n < p, the OLS solution is not even unique (perfect multicollinearity)

4. You use p and k as the number of variables simultaneously.
5. The name of the paper does not fully correspond to its meaning.

Questions

1. What is the difference between multicollinearity and net collinearity? Bring an example.

2. Can we use another loss function? Why do we use square?

Our estimates of the population parameters are referred to as $\hat{\beta}$. Recall that the criteria we use for obtaining our estimates is to find the estimator $\hat{\beta}$ that minimizes the sum of squared residuals. Why this criteria? Where does this criteria come from?

3. Does the Lasso have the b coefficient?

4. Var (ols) ><? Var(ridge), bias(ols)<>?bias(ridge)

Summary

Criteria	%	Points	Anna	Anush
Curiosity	20%	0.8	0.7	0.7
Skepticism	35%	1.4	0.8	0.8
Organization	25%	1	0.8	0.8
Shiny	20%	0.8	*	*
Total	100%	4	2.3	2.3

YSU FEM DSB Final Project Feedback - 1002/U91 |

Mary, Sona

Coding

1. I could not find anything new in your code

Conceptual

1. I could not find anything new in your report
2. A lot of typos which make your work less attractive.
3. You cannot classify the price:

   In this paper we have used KNN algorithm to classify Mobile phone prices depending on i

4. The existence of incorrect statements e.g.:

   Classifiers Of Machine Learing:

   1. Decision Trees
   2. Bayesian Classifiers
   3. Neural Networks
   4. K-Nearest Neighbour
   5. Support Vector Machines
   6. Linear Regression
   7. Logistic Regression

5. The usage of standardization for categorical data is not correct.

```
standardized.X <- scale(ds[,-21])
set.seed(55)
training.index <- createDataPartition(
```

Questions

1. How the accuracy, sensitivity, specificity for multiple classes are calculated?
2. How the Minkowski distance is calculated?
3. What does happen with DB with an increase of k?

Suggestions, Summary

You can play on train-test errors, decision boundaries, the problem of high dimensionality, similarity, dissimilarity measures.

Criteria	%	Points	Mary	Sona
Curiosity	20%	0.8	0.6	0.6
Skepticism	35%	1.4	0.8	0.8
Organization	25%	1	0.7	0.7
Shiny	20%	0.8	0.8	0.8
Total	100%	4	2.9	2.9

YSU FEM DSB Final Project Feedback - 1002/U91 |

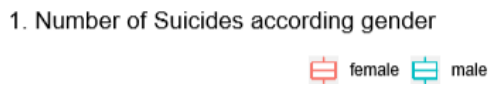
Hripsime, Anjel

Coding

1. I could not find anything new in your code.

Conceptual

1. A lot of typos, copy-past formulas, the difference in fonts, scientific notations which make your work less attractive.
2. The labeling of graphs and variables are not correct:



3. The usage of boxplots for high imbalance (with a high frequency of 0-s) data is not appropriate.
4. The inclusion of year in regression should be done by data manipulation.
5. The interpretation of significance and coefficients (both for nominal and ordinal data) are wrong.

…for a one unit change in the age75+ years male, the difference in the logs of expected counts of the sum_Suicide is expected to decrease…

…We can see that for a one unit change in the generation^4…

6. The testing of overdispersion was needed to understand the estimation method.
7. The usage of rounding in the report is necessary.

Questions

1. What is the shape of Poisson regression fitted line?
2. How did you select the variables?
3. How do we estimate the Poisson regression?
4. What does Poisson heterogeneity mean?
5. How did you conclude that var is greater then mean?

$$\mathrm{Var}[Y_i|x_i] = \exp(\beta^t x_i)(1 + \eta^2 \exp(\beta^t x_i)).$$

Summary

Criteria	%	Points	Anjel	Hripsime
Curiosity	20%	0.8	0.7	0.7
Skepticism	35%	1.4	0.5	0.5
Organization	25%	1	0.7	0.8
Shiny	20%	0.8	*	*
Total	100%	4	1.9	2

YSU FEM DSB Final Project Feedback - 1002/U91 |

Elen, Anna

Coding

1. The absence of source working files (Rmd).

2. Path Error:Level:God

```
offers <- read.csv('C:/Users/chilinga/Desktop/master_degree/data_mining/Project/Offers.csv', sep = ';', header=T)
head(offers)

trans <- read.csv('C:/Users/chilinga/Desktop/master_degree/data_mining/Project/Transaction.csv', sep = ';', header=T)
head(trans)
```

3. Libraries must be in the first chunk:
4.

```
{r}

library(tidyverse)
library(reshape2)
library(plyr)
library(pivottabler)

offers <- read.csv('Offers.csv', sep = ';', header=T)
head(offers)

trans <- read.csv('Transaction.csv', sep = ';', header=T)
head(trans)
library(reshape)
```

5. The absence of cleaned data:

```
trans <- read.csv('Transaction.csv', sep = ';', header=T)
head(trans)
```

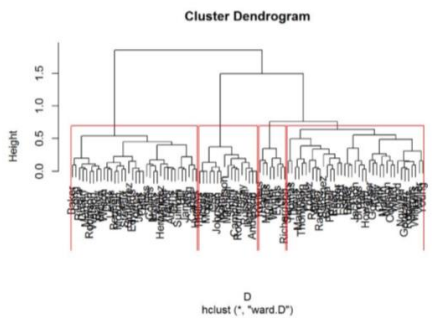
	i...Offer.. <fctr>
1	Smith,2
2	Smith,24
3	Johnson,17
4	Johnson,24
5	Johnson,26
6	Williams,18

6 rows

```
offers
```

Offer...Campaign.Varietal.Minimum.Qty..kg..Discount.....O <fctr>
1,January,Malbec,72,56,France,FALSE
2,January,Pinot Noir,72,17,France,FALSE
3,February,Espumante,144,32,Oregon,TRUE
4,February,Champagne,72,48,France,TRUE
5,February,Cabernet Sauvignon,144,44,New Zealand,TRUE
6,March,Prosecco,144,86,Chile,FALSE
7,March,Prosecco,6,40,Australia,TRUE
8,March,Espumante,6,45,South Africa,FALSE
9,April,Chardonnay,144,57,Chile,FALSE
10,April,Prosecco,72,52,California,FALSE

6. Visualization should help to understand the data/tool.



Conceptual

1. We can perform clustering with different measures.

1. Համոզվել, որ տվյալներն ամբողջական են, և չկան բացակայող տվյալներ, բնութագրիչներն ունեն չափման նույն միավորը:

2. The discordance in text and work. You do not use the single-link:

Գոյություն ունի կլաստերինգի 2 հիմնական տեսակ՝ հիերարխիկ և k-միջիններ:
Մեր օրինակում դիտարկել ենք հիերարխիկ տեսակը single-link տարբերակով:

3. The wrong/absence of/ description of variables (offer, past peak).
4. Ambiguous statements:

Ինչպես գիտենք, մի կլաստերի ներսում հաճախորդները իրար նման են, միևնույն ժամանակ տարբեր են մյուս կլաստերների հաճախորդներից: Այժմ հաշվենք, թե յուրաքանչյուր հաճախորդը որքան է տարբերվում կլաստերի միջինից: Այս հեռավորությունը կարող ենք հաշվել տարբեր ինդեքսներով. այդ դեպքում մենք հաշվել ենք Gower-ի տարբերությամբ:

5. It is impossible to obtain overlapped clusters from hierarchical clustering.
6. The linkages are used for distance measures of clusters (not points).
7. The absence of scaling.

Questions

1. Why did you choose k=4?
2. What did the function daisy do?
3. How does the method Gower work?
4. Why did you use hierarchical clustering?
5. What does method = “ward” mean?
6. What is the difference between within cluster distance, total cluster distance, between cluster distance?

Summary

Criteria	%	Points	Elen	Anna
Curiosity	20%	0.8	0.78	0.78
Skepticism	35%	1.4	0.6	0.55
Organization	25%	1	0.6	0.6
Shiny	20%	0.8	*	*
Total	100%	4	2	1.9

Tigran

Coding

1. The absence of RMD file.
2. Path Error:Level:God

```
setwd("C:\\Users\\Lenovo\\Desktop\\Final Project")
getwd()
```

Conceptual

1. Weak structure of report (into-model-body-conclusion), lack of continuity/integrity:

   Deteils in Presentation. Should speak about lambda parameter.

2. .. the predicted outcome is the class (discrete).. is not the same.
3. You do not need to have the long R outputs in your main report.

Questions

1. What is the difference between bagging and random forest?
2. What is the difference between the validation set ans test set?
3. Why do we need regularization term in XGBoost?
4. What does weight mean in XGBoost?
5. What is the parameter T in regularization term?

Summary

Criteria	%	Points	Tigran
Curiosity	20%	0.8	0.8
Skepticism	35%	1.4	0.65
Organization	25%	1	0.86
Shiny	20%	0.8	0.8
Total	100%	4	3.1