MGT7179- Assignment 2

Mohsen Jafari

2024-01-11

Introduction

In this assignment, similar to the previous ones, each student receives a customised dataset. The dataset allocated to each student will be shared via Canvas (***Please check Canvas page regularly, and put its update to automatic***).

Online reviews have become a popular and influential source of information for consumers who are looking for various products and services. Moreover, online reviews can provide potential customers with insights into the quality, performance, and reputation of the different services provided by firms, as well as their personal experiences and satisfaction with the obtained services. However, online reviews can also pose some challenges and limitations, such as the credibility, validity, and reliability, and the difficulty of measuring and comparing the complex and multidimensional aspects of service quality and customer satisfaction.

In this assignment, you will use a dataset that contains information about some online and the reviews they received on multiple platforms. You need to use a customized portion of the given dataset which includes variables such as field, state, CEO graduation year, CEO graduation school, CEO gender, total number of reviews, rating.

The goal of this assignment is to help you develop your skills and knowledge in statistical learning, and to apply them to a real-world dataset that has practical and social implications. You will also learn how to interpret and visualize the results of your models, and how to understand the advantages and limitations of each model. You will also learn how to critically evaluate the usefulness and appropriateness of online reviews of doctors, and how to identify and address the challenges and issues that arise from them.

This assignment count for 60% of the overall mark.

Dataset and its structure

This dataset contains online reveiw and other characteristics of firms . In addition, most of data columns are categorical. More precisely, among all of the ***twenty*** data columns of the dataset (see below), mainly, the following ones are non-categorical variables: **score, review_count, Tot_Clms_Services, Br_Tot_Clms_Services, Gnrc_Tot_Clms_Services, Othr_Tot_Clms_Services, Opi_Tot_Clms_Services, Ant_Tot_Clms_Services** Thus, you ***must*** note that most of the other variables, are categorical even though they have integer values. For example, ***CEO_sch_cat*** and ***field_cat*** are categorical variables, and have than multiple levels.

The data columns (data dictionary) of the data set is shown below:

```
## |Original|Description|
## |------------------------|------------------------|
## |Platform|Platform from which reviews extracted|
## |business_ID|ID allocated to a business|
## |city|City in which business operates|
## |state|State in which business operates|
## |postal_code|Postal code of business office|
## |score|Total review score on platform|
## |review_count|Total number of reviews|
## |Gender|Gender of CEO|
## |CEO_sch_cat|School CEO attended|
## |CEO_Grd_yr|Graduation year of CEO|
## |field_cat|Field related to the type of provided service|
## |ZIP Code|ZIP code of business|
## |Business_ID_other|Another ID of business|
## |Rural_metropolitan_Desc|Description of the area business located|
## |Tot_Clms_Services|Total number services 1|
## |Br_Tot_Clms_Services|Total number services 2|
## |Gnrc_Tot_Clms_Services|Total number services 3|
## |Othr_Tot_Clms_Services|Total number services 4|
## |Opi_Tot_Clms_Services|Total number services 5|
## |Ant_Tot_Clms_Services|Total number services 6|
```

Initial Analysis

- Item A1: As each student is allocated a subset of ***Platforms, or field_cat, or state***, he, or she needs to load the given dataset as a dataframe, and filter the given csv file (will be announced on Canvas), and create a dataset which is the one related to him/her (4 marks),.. **Please note that which subset of the variables you need to use to build your customized dataset will be announced on Canvas**…
- Item A2: Each student needs to proivde a summary of all the blank,zero value records from the given dataset A, and remove them, and create another dataset B too, and compare the two datasets.. (6 marks).
- Item A3: As highlighted earlier, there are categorical variables in the dataset. Thus, each student should code or covert the columns of the given dataset into qualitative or categorical variables, if necessary (as.factor in R); For example, and for ***emphasizing again***, in the code chunks below, I have created dummy variables for two data columns: (i) `CEO_sch_cat'', and (ii)` field_cat''. As this is critical, and by taking this, ***total number of predictors*** will be about ***a large number depending on the allocated subset to a student***, its completion provides 12 marks for a student.
    - Remark: Please note that the only target,response variable among the categorical data columns is ***score***. Also, **Please examine both pdf and html version of the assignment, as with the html version, it is possible to scroll horizontally and fully examine the dataset tables**.

```
#tinytex:::install_prebuilt()
#install.packages("blogdown")

# install.packages("htmltools")
# devtools::install_github("kupietz/kableExtra")
# blogdown::build_site(build_rmd = T, local = T)
df0 <- read.csv("student_merge_platform_business_file_final05_edited_selectedRows.csv")
# install.packages("systemfonts")
# Load the library
library(fastDummies)
# df0$Proudct.issue.type
# install.packages('kableExtra')
library(kableExtra)
results <- fastDummies::dummy_cols(df0, select_columns = "CEO_sch_cat")
#kable(results, format="latex", booktabs=TRUE) %>%
#  kable_styling(latex_options="scale_down")
#knitr::kable(results)
# kable(results) %>%
# kable_styling(font_size = 3)

kable(results) %>%
  kable_styling(font_size = 7)
```

X	platform	business_ID	city	state	postal_code	score	review_count	Gender	CEO_sch_cat	CEO_Grd_yr	field_cat	ZIPCode	Business_ID_other	Rural_metropolitan_Desc	Tot_Clms_Services	Br_Tot_Clms_Services	Gnrc
0	platform 1	1770720401	Santa Barbara	CA	93101	5.0	7	F	108	1997	13	931032109	NA		NA	NA	
1	platform 1	1699268318	Clearwater	FL	33755	5.0	10	F	108	2017	21	337631726	1699268318	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	971	138	
2	platform 1	1336331669	Bala Cynwyd	PA	19004	4.0	13	M	108	1986	12	190043207	NA		NA	NA	
3	platform 1	1528008463	Plymouth Meeting	PA	19462	2.5	8	M	138	1980	75	194621718	1528008463	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	1988	NA	
4	platform 1	1366867269	Voorhees	NJ	8043	3.5	17	F	116	2014	69	80434509	1366867269	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	847	73	
5	platform 1	1689735383	Tarpon Springs	FL	34689	2.0	29	M	170	1988	55	346893790	NA		NA	NA	
6	platform 1	1356471593	Tampa	FL	33607	3.0	36	M	116	1992	21	805162422	1356471593	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	869	103	
7	platform 1	1023598968	Brownsburg	IN	46112	1.5	14	M	108	2018	58	461121031	NA		NA	NA	
8	platform 1	1316037153	Tucson	AZ	85718	2.5	43	M	3	1992	33	104672401	NA		NA	NA	
9	platform 1	1902048259	Saint Petersburg	FL	33713	1.5	6	M	201	2009	3	337051300	1902048259	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	24	11	
10	platform 1	1821040866	Indianapolis	IN	46203	2.5	10	M	210	1998	18	462378606	NA		NA	NA	
11	platform 1	1578029260	Reno	NV	89512	4.5	19	F	108	2016	66	895570001	NA		NA	NA	
12	platform 1	1770652448	Reno	NV	89509	4.0	10	F	108	2012	21	895113038	1770652448	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	60	NA	
13	platform 1	1124065081	Langhorne	PA	19047	3.0	7	F	51	1985	3	190531556	NA		NA	NA	
14	platform 1	1700221603	Voorhees	NJ	8043	4.5	6	M	108	2013	33	80439612	1700221603	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	246	22	
15	platform 1	1124201801	Reno	NV	89521	4.0	7	M	108	2007	63	895021463	1124201801	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	358	33	
16	platform 1	1821040866	Philadelphia	PA	19128	3.0	11	M	210	1998	18	191115729	NA		NA	NA	
17	platform 1	1265069702	Tucson	AZ	85715	2.0	15	M	1	2020	19	757012036	1265069702	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	66	NA	
18	platform 1	1497909147	Philadelphia	PA	19102	3.5	120	F	213	1988	63	191074315	NA		NA	NA	
19	platform 1	1477798676	Tampa	FL	33609	4.0	55	M	204	2007	64	752082312	NA		NA	NA	

```
# kable(results, "html") %>% kable_styling("striped") %>% scroll_box(width = "100%")
```

```
results <- fastDummies::dummy_cols(df0, select_columns = "field_cat")
kable(results) %>%
  kable_styling(font_size = 7)
```

X	platform	business_ID	city	state	postal_code	score	review_count	Gender	CEO_sch_cat	CEO_Grd_yr	field_cat	ZIP.Code	Business_ID_other	Rural_metropolitan_Desc	Tot_Clms_Services	Br_Tot_Clms_Services	Gnrc
0	platform 1	1770720401	Santa Barbara	CA	93101	5.0	7	F	108	1997	13	931032109	NA		NA	NA	
1	platform 1	1699268318	Clearwater	FL	33755	5.0	10	F	108	2017	21	337631726	1699268318	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	971	138	
2	platform 1	1336331669	Bala Cynwyd	PA	19004	4.0	13	M	108	1986	12	190043207	NA		NA	NA	
3	platform 1	1528008463	Plymouth Meeting	PA	19462	2.5	8	M	138	1980	75	194621718	1528008463	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	1988	NA	
4	platform 1	1366867269	Voorhees	NJ	8043	3.5	17	F	116	2014	69	80434509	1366867269	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	847	73	
5	platform 1	1689735383	Tarpon Springs	FL	34689	2.0	29	M	170	1988	55	346893790	NA		NA	NA	
6	platform 1	1356471593	Tampa	FL	33607	3.0	36	M	116	1992	21	805162422	1356471593	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	869	103	
7	platform 1	1023598968	Brownsburg	IN	46112	1.5	14	M	108	2018	58	461121031	NA		NA	NA	
8	platform 1	1316037153	Tucson	AZ	85718	2.5	43	M	3	1992	33	104672401	NA		NA	NA	
9	platform 1	1902048259	Saint Petersburg	FL	33713	1.5	6	M	201	2009	3	337051300	1902048259	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	24	11	
10	platform 1	1821040866	Indianapolis	IN	46203	2.5	10	M	210	1998	18	462378606	NA		NA	NA	
11	platform 1	1578029260	Reno	NV	89512	4.5	19	F	108	2016	66	895570001	NA		NA	NA	
12	platform 1	1770652448	Reno	NV	89509	4.0	10	F	108	2012	21	895113038	1770652448	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	60	NA	
13	platform 1	1124065081	Langhorne	PA	19047	3.0	7	F	51	1985	3	190531556	NA		NA	NA	
14	platform 1	1700221603	Voorhees	NJ	8043	4.5	6	M	108	2013	33	80439612	1700221603	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	246	22	
15	platform 1	1124201801	Reno	NV	89521	4.0	7	M	108	2007	63	895021463	1124201801	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	358	33	
16	platform 1	1821040866	Philadelphia	PA	19128	3.0	11	M	210	1998	18	191115729	NA		NA	NA	
17	platform 1	1265069702	Tucson	AZ	85715	2.0	15	M	1	2020	19	757012036	1265069702	Metropolitan area core: primary flow within an urbanized area of 50,000 and greater	66	NA	
18	platform 1	1497909147	Philadelphia	PA	19102	3.5	120	F	213	1988	63	191074315	NA		NA	NA	
19	platform 1	1477798676	Tampa	FL	33609	4.0	55	M	204	2007	64	752082312	NA		NA	NA	

- Item A4: Each student should investigate the predictors graphically, using scatter plots or other tools, highlighting the relationships among the predictors (15 marks),
    - Remark: Since most probably, there are a large number predictors, for this item, each student should select 3-4 fields (e.g., field_cat=3, 13, 21), and for each, identify the top 10-20 most frequent predictors (e.g., predictors related to 'score'), and finally, generate the relevant plots, or apply other relevant tools.
    - Each student should write 1-2 short paragraphs explaning the insights he/she finds for this item (as well as any logic, even arbitrary, used to select the three fields). ***Further details*** should be provided in the Appendix section.

Statistical Learning Models

Each student may take the following steps in order to understand how predictors (both categorical & non-categorical) are associated with the response variable, ***score*** (Y="score").

- Item B1: Each student should select one field (e.g., field_cat=3) among the given fields to him/her in the dataset (announced on Canvas). Then, and in the next step, in order to prevent over-fitting and complexity in the presence of a large number of levels in the data columns, he/she should run a ***Lasso regression*** model to select the top, critical predictors (mark 15).
    - Remark I: If a student cannot run a Lasso model, he/she may use other methods like subset selection algorithms (see chapter 6 of the core book, "Linear Model Selection and Regulariation"), however, it should be noted that those algorithms are highly computationally expensive, and thus, a student may repeat subset selection 4-5 times each time with a subset of ***only ten*** predictors. In the case of applying subset selection algorithms, the maximum possible mark will be 7 out of 12 marks (student loses 6 mark).
    - Remark II: Use cross-validation to choose the optimal values of the regularization parameters, and also, compare the coefficients of the models and identify which variables are shrunk or set to zero by lasso..
    - **Note**: Each student should repeat item B1 for at least ***three*** different fields (e.g., field_cat=15), and create a summary table (or list) showing the differences between subsets of critical predictors identified ***across fields*** (e.g. for the field coded as 6, the set of critical predictors were X_4, X_10, X90, whereas, that set for the field coded 5 includes X_1, X_20,..). The mark for this task is (9 marks). Provide a short paragraph of what you find, and ***further details*** (if exist) should be provided in the Appendix section.

- Item B2: Each student, for each of ***the four fields in the dataset Item B1**, should develop a Generalized Additive Model (GAM) to predict score of an online business given the identified predictors in Item B1 (20 marks)
    - In other words, each student should use each of the four fields. Next, he/she filters the data records related to that field, and then, finally, develops a GAM model for the resulting dataset.
    - Students need to plot the partial residuals and interpret the results.
- Item B3: Each student, for each of ***the four fields in Item B1**, should develop one statistical model (e.g., decision tree, support vector machine, deep learning) to predict score of an online business given the identified predictors in Item B1 (20 marks)
    - In other words, each student should use each of the four fields, filters the data records related to that field, and then, develops only one particular model (***ONLY one model type, and not more than one model types***) for the resulting dataset.
    - Remark I: The used model in Item B3 must be among those models covered after the 5th lecture (All materials in the chapter7 and subsequent chapters of the core book of the module covered during the lectures-**An introduction to statistical learning**..);
    - Remark II: In the case of aiming to apply classification model (e.g., SVM, classification tree), you may convert the response variable (review score) into a binary variable based on low ($score \in [0-3]$) and high ($score \in (3-5]$) score values, and elaborate classes..
    - Remark III: Use cross-validation to choose the optimal values of parameters(or tune hyperparameters) for models..
- Item B4: How accurate are the results of items B2 & B3? Which of these approaches yields the best performance? Compare the performance of the model using possible measures like adjusted R-squared, AIC, BIC, RMSE, and MAE.., or bias-variance trade-off, interpretability .. (8 marks)

Submission and deadlines

- You need to write your report in a Rmarkdown file and generate a pdf file from it, and ***submit both of the files*** (if you cannot generate pdf file from our Rmarkdown file, generate an html file, and print that as a pdf file). Furthermore, your report structure can be similar to the one for the first assignment too (e.g., Introductions, methods, model,.. discussion).
- The link of the dataset for each student will be provided on Canvas,
- Deadline for submission of reports & codes is Monday, **May 06, 2024, 23:59**; Students are expected to submit on the stated deadline unless they submit an EC.