Project Instructions

Business Decision Algorithms*

FINAL SUBMISSION DATE = 10/26/2023 (11:59 pm MST)

Score = out of 50

No late submissions will be accepted.

Students can submit the project anytime during the 3-week window.

Please read Project FAQs on Canvas carefully.

Submission Guidelines

1. **Project submissions can be group submissions with groups not exceeding 3 students per group.** Students are free to submit individually if they prefer. Groups will not be formed for students, they can form groups on their own or choose to submit individually. **Inform the instructors by 10/12/2023 whether you plan to submit individually or in a group. Copy all group members in the email to ensure everyone is on the same page. If we don't receive any information from a student, we will assume they are submitting individually.**

2. The project submission will be responses to the **8 questions** to be submitted on Canvas using the Assignment link **Project Submission**. Submission should be in R Markdown file (the template posted on Canvas can be edited for submission) with the name of all student(s) in the group. While grading we should be able to run the R code and reproduce the results. Submissions will lose points if the code does not run.

3. **No extension of deadline will be provided** except for an emergency approved by the office of accommodations. After 11:59 pm on 10/26/23, students will NOT be able to submit their project and will lose 50% of their course grade.

Academic honesty

Turnitin and other digital checks will be used to check for academic honesty for all submissions. Please do not share your responses with other students. The response should be the students' original responses and not copied from elsewhere. Peculiarities and similarities in project submission to current and past submissions, will be flagged and investigated. Do not post or share any course material without the instructor's permission.

Project Submission Instructions

1. There are two data files (one for the different regressions and the other for text analysis). Hint code to read in the data and perform analysis has been provided in this document.
2. Students are encouraged to improvise the class R code (available in class notes) and ***troubleshoot the code on their own***. As mentioned in the syllabus, this is the responsibility of the students.
3. If you get error messages, it might be a localized problem on your machine. Please trouble shoot.
4. Set seed to ensure consistently reproducible output.
5. To maintain consistency in grading, **we cannot go over and check project submission drafts before the final submission**. You are free to ask specific questions. Please post project-related questions in the final project discussion rather than emailing instructors. Please also review the final project discussion board prior to posting to ensure your question has not already been asked.
6. Students will need to infer the results from the analysis and answer the questions. Hence, being able to interpret the results from the analysis is required. Clear and detailed justification should be provided for each proposed business strategy in the student's response. Use strategies, business principles, methods discussed in class. **Unclear answers or lack of justification will result in loss of points.**

PROJECT DATASET

Regression dataset: http://data.mishra.us/files/project_data.csv
Text analysis dataset: http://data.mishra.us/files/project_reviews.csv

PROJECT CASE

A movie chain in the southwest region, MovieMagic is considering ways in which it can increase spending on concessions. It has collected information of 2000 of its customers, some of whom are part of their loyalty program and some who are not. They have information on the following 8 variables, which they plan to use as predictors. They plan to use *amount_spent* (i.e. the amount spent on concessions) as the outcome variable since they have learnt from observations that much of the profit is derived from concession sales.

Predictors (note: some of these variables are categorical i.e. this needs to be addressed in the R code, else the analysis won't run)

age = age of the customer

job = type of job e.g. management, technician, entrepreneur,

streaming = how many streaming services is being subscribed to

education = primary, secondary, tertiary

seen_alone = whether the movie was seen alone or with some others (yes/no)

discount = whether they were sent a discount coupon (yes/no)

days_member = days since member of MovieMagic

movies_seen = number of movies seen last time period

Outcome

amount_spent = amount spent on concessions

Their objective is to find out what factors can increase concession sales and how they can improve their prediction of the outcome variable so that they can plan better.

Along with amount_spent, MovieMagic was also able to collect information from about 75 of its existing customers in the form of reviews. They feel that this text data can provide a different insight into what customers think and feel about MovieMagic.

They realize that their objective has two components: interpretation and prediction. Hence, they decide to run 3 different types of analysis.

1. Linear regression
2. Penalized Regression
3. Text analysis

Consider that you have been asked to run the analysis and answer the following questions MovieMagic wants answered.

Project Questions

These are the questions that MovieMagic wants answers to:
For questions 1, 2, 3, 4, and 5, the first dataset **project_data** would be used.

1. Of the 8 predictors, which predictors have a significant influence on amount spent on concessions? Which predictors are multicollinear? Justify your response with reasons from the analysis.

2. Which predictors have a positive influence and which predictors have a negative influence on the amount spent on concessions? Which analysis, regression or penalized regression, helped you answer this question? If you ran a neural net model, can it help you find the significant (or not) predictors and their magnitude and direction of influence on the outcome?

3. Which analysis, linear regression or penalized regression, helps you select relevant variables? Which predictor variables would you use in the model? Justify your answer using the analysis. Would a Ridge or a LASSO help in selecting relevant predictors?

4. If you split the data 70-30 versus 80-20, how does it influence RMSE and R-squared values of the linear regression?

5. Given the regression analysis, what strategies can MovieMagic come up with to increase amount spent on concessions? Discuss the magnitude and direction of the anticipated effect. Use both statistical justification and a simplified explanation (anticipating many decision-makers at MovieMagic may not know all the technical jargon).

For questions 6 and 7, students would be using the second dataset, the **project_reviews** dataset.

6. MovieMagic wants to visualize the reviews through a *wordcloud* and wants to find out which words are used most commonly in the reviews that customers write for MovieMagic. Create 2 wordclouds - one for reviews that received 3, 4, or 5 star ratings and another with reviews that received 1 or 2 stars ratings. Knowing the prominent words in each of the wordclouds, what strategies can be developed in messaging customers? Would the strategies differ?

7. MovieMagic also wants to use topic modeling to find out whether the content in the reviews could be categorized into specific topics. If you used LDA to create 3 topic groups (k = 3), MovieMagic wants you to use the words within the 3 topics to infer **topic title**. Which term is the most relevant in each of the three topics and how would it inform your business strategy? Given the topics you inferred what strategies would you suggest are possible for MovieMagic if it wants to increase concession sales. Would you recommend promotions or advertising or loyalty program; justify your choice of business strategy?

Using insights from both datasets answer question 8

8. MovieMagic asks you whether your analysis reflects a causal relationship. Discuss any limitations of the dataset and your analysis regarding causal inference. What experiment might you recommend given these limitations and your analysis? What would be the experimental design? How would this lead to a deeper understanding of what business strategies would work? Make sure to clearly define the input variables, main effects, interactive influences (if any that you want to test for) and the outcome variable. *Example – using the top terms from the LDA a 3 cell experiment can be designed to find out how using these terms in messaging before the movie begins influences concession sales.

CODE HINTS

Here are a few hints for the project. Students are encouraged to try codes of their own before using this provided code. The aim of the project is to help students get comfortable with conducting analysis on their own for which the R-code for each class has been provided on RStudio-cloud as well as the output described in the class notes.

The provided code hint is ONE way of performing the analysis. You should explore other ways too.

Regression analysis

First read in the data. You can execute the code in R only when you have read in the data. Install the required libraries. Some libraries are indicated here. But depending on the R package you plan to use for the analysis this would change.

```
# install libraries

library(dplyr)

# Read in the data
data1 <- read.csv("http://data.mishra.us/files/project_data.csv")
```

Linear Regression

Performing a linear regression on the entire data. Hint: the code to perform the linear regression is provided here.

```
model1<- lm(amount_spent~.,   data=data1)
summary(model1) # will give output for each level of each categorical predictor
```

Testing for multicollinearity Note: some of these hint codes and questions that appear with the code are provided to help you think about the analysis and inferences you can obtain from the analysis. You do not have to answer these questions. The submission should include answers to the **8 Questions** provided earlier. What can be inferred about multicollinearity in the data? What threshold value needs to be kept in mind for determining whether a variable has high multicollinearity or not? Which predictor variable has high multicollinearity according to the VIF from the output?

```
library(car)
vif(model1)
```

Which predictors are most relevant for understanding their influence on the outcome variable?

```
set.seed(149)
library(glmnet)

#Selecting relevant predictors using penalized regression


#LASSO


#Ridge
```

Building a predictive model Split data into train and test sets. Right now the split is train = 70% and test set = 30%. What would happen if we used the entire data set for running the linear regression model without splitting it into train and test set? What would happen to the analysis, results, and business insights if the train-test split is changed to 80% (train) and 20% (test)?

This is one way of splitting the data. Students are free to try other methods

```
library(caret)
set.seed(1234)
datasplit <- createDataPartition(data1$amount_spent, p = 0.7, list=FALSE)
trainData <- data1[datasplit,]
testData  <- data1[-datasplit,]
```

Run a prediction model and see what is the RMSE and R-squared values. What is the R-squared for the linear regression model?

```
#training the model

# Obtaining predictions from the model

#Calculating RMSE, R-squared, and MAE for the predicted model values
```

Text Analysis

Read in the reviews data from the code given next. Then use the relevant libraries for performing text analysis. The class code for creating a *wordcloud* and for running a topic model will help you perform this text analysis.

Edit code as needed

Students are free to use other packages and libraries for the analysis

```
library(tidyverse)
library(topicmodels)
library(ggplot2)
library(dplyr)
library(tidytext)
library(tidyr)
library(RTextTools)
library(wordcloud)
library(tm)
library(stringr)
library(quanteda)
library(reshape2)
library(quanteda.textplots)

text <- read.csv(url("http://data.mishra.us/files/project_reviews.csv"))

# categorize rating into two groups

text$valence[text$star == 1 | text$star == 2 ] = "Negative"
text$valence[text$star == 3 | text$star == 4 | text$star == 5     ] = "Positive"
text$text <- as.character(text$text)
set.seed(1234)

#create a wordcloud based on the positive versus negative valence
```

Topic Model

Now run the analysis for obtaining **topic models**. Read the project question carefully. How many topics need to be produced for analysis?

Edit code as needed

```r
# perform a Latent Dirichlet Analysis
text <- read.csv("http://data.mishra.us/files/project_reviews.csv")
# first remove stop words
corpus <- VCorpus(VectorSource(text$text))
# a function to clean /,@,\\,|
toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
corpus <- tm_map(corpus, toSpace, "/|@|\\|")
corpus<- tm_map(corpus, stripWhitespace) # remove white space
# covert all to lower case else same word as lower and uppercase will classified as different
corpus <- tm_map(corpus, content_transformer(tolower))
```

```r
corpus <- tm_map(corpus, removeNumbers) # remove numbers
corpus <- tm_map(corpus, removePunctuation) # remove punctuations
corpus <- tm_map(corpus, removeWords, stopwords("en"))

dtm <- DocumentTermMatrix(corpus)

set.seed(234)
rowTotals <- apply(dtm , 1, sum)
dtm   <- dtm[rowTotals> 0, ]
lda <- LDA(dtm, k = 3, method = "Gibbs", control = NULL)
topics <- tidy(lda, matrix = "beta") # beta is the topic-word density
```