MSDS 6372 Project 1 Details

Introduction

This data analysis and presentation is your opportunity to apply many of the tools and modeling strategies discussed in the course so far. Students have been assigned to groups by Dr. Turner. Each group may select one of the two data sets listed in the Appendix. If another data set is to be used, the group must discuss with me for final approval no later than Unit 4 live session. For each data set there will be two main objectives to address. The first objective is very specific to the data set assigned. This objective is meant to be unique to allow students to think about what approaches, procedures, and methods, should be used to answer the main question. This objective may seem somewhat vague and that is intentional. In most real-world problems, people do not give direction on what needs to be done. It’s up to us to assess the situation, plan an approach, and successfully communicate the results from that approach to answer the main questions the study set out to do.

If you have follow up questions, please reach out to me so we can discuss. Talking back and forth with other people related to the data is also a skill worth improving. The second objective is more direct and applies to essentially to both data sets. Each of the data sets objectives are available in the Appendix as well along with a brief data description.

Project Deliverables and Due Date

Only one person needs to submit the project by Feb 9th at 11:59pm. In summary there should be 3 items in the submission.

1. Presenation file
2. Video recording of presentation (youtube, mp4, etc)
3. R file of work

Each group will present their analysis in a 20 minutes presentation. This time requirement WILL feel tough, but it is important to spend enough time to strategize with your group on a way to effectively communicate the most critical aspects of your report. This strategy should be deliberate and thought out over the coming weeks. All group members are expected to participate during the presentation.

In addition to the presentation video, each group should provide the ppt slides as well as an R script (or notebook) of the code that produced the analysis presented. For things that didn’t make it into the presentation but you feel are important, you may cite your R script for me to look at in further detail or provide those details at the end of the ppt slides for me to review.

Lastly, group members will be asked to fill out a peer review sheet as a way for me to gauge each members engagement and participation. This should not be submitted with the project. There is an individual assignment within CANVAS for each individual to submit their peer review. The peer review form will be provided in the FILES folder.

Grading Rubric

As the objectives are slightly different for each group, it is hard to define a single grading rubric when it comes to conducting specific analysis. At a high level a general point breakdown of the presentation will be something like the following:

|  |  |
| --- | --- |
| Item | Points |
| Introductory Material | 10 |
| EDA | 20 |
| Addressing Objective 1 | 20 |
| Addressing Objective 2 | 20 |
| Final Remarks | 10 |
| Overall organization and effective communication and peer review. | 20 |

**Required Information and SAMPLE FORMAT**

Introduction and Objective Summary

Data Description / Processing Summary

Must include a table of variable names

Must include summary statistics

Exploratory Data Analysis

Minimum of 4-6 graphics showing potential predictor trends (or lack there of) with the response. At least 2 must demonstrate an exploration of potential interaction terms

Objective 1: *Must include a high level explanation of the model fitting approach.*

*Which variables were included/exclude along with how/why?*

*Feature Selection Summary if applicable*

*The final model should be clearly defined.*

*Summary table of coefficients*

Testing and Confidence interval results to address the objective’s key question(s)

There should be at least one formal interpretation of a regression coefficient along with a confidence interval. For some objectives, additional interpretations or summaries will be needed. It is up to the group to decide on what is appropriate.

Objective 2: Summary of approach to include complexity in the regression model

Perhaps additional EDA to motivate (situational / those interaction plots required for EDA may be brought up here as well if it makes sense)

Model comparison table and recommendation

Conclusion: Quick summary of the findings in objective 1 and 2.

Additional discussion points if applicable/time permitted: Scope of inference? What would you do if given more time? Recommendations moving forward? Insight the model gave? Etc.

Appendix: Do not need to present but if you think there are additional tables or graphics worthy of reference, then place them in the back and mention them while presenting or reference the .RMD file.

Things to think about

Introductory Material should cover the main objectives of your project as well as a data description table (like the ones provide in the appendix). A brief outlier of what you’re going to discuss can be helpful if it makes sense.

EDA should include basic summary statistics of each variable. Additional graphics such as scatterplots, boxplots, etc should be considered to help the audience understand what trends appear to exist with the response and/or to address muticollinearity concerns.

Objective 1 should clearly summarize the general approach to how you obtained your final model used to answer the question. This should also include an assessment of residual diagnostics and addressing/commenting on any concerns when it comes to outliers, influential points, or assumptions of the MLR. Any tests used should be formally communicated and result tables should discuss the general interpretation of the coefficients from your model along with confidence intervals. You do not need to interpret every predictor in the model as your model may be quite large. Groups should consider carefully what tests and/or confidence intervals could be used to help answer the objectives question(s). There should be at least one formal interpretation of a regression coefficient along with a confidence interval. For some objectives, additional interpretations or summaries will be needed. It is up to the group to decide on what is appropriate.

Objective 2 is more open for you to explore more complicated models without the need for interpretation. Consider including EDA graphics that led you to consider the complexity that you tried. It can help us make sense if the added complexity is really that helpful or not.

Final remarks should consider discussion involving the population (scope of inference). Discussing potential issues that the team couldn’t overcome or what you would have liked to do if you had more time are all welcome discussions if it makes sense to include or time allows.

Overall organization and effective communication. The biggest advice I can give is do not let the graphics do the talking for you. A good mental exercise is to ask yourself if you discussed that verbally? If the answer is no assume that your audience doesn’t know about it then. Do not be afraid to isolate output. If you have a huge table of results but you only want to talk about one component of the output, then just provide that or highlight it so everyone is aware of what you are focusing on. See Unit 4 for additional discussions and ideas.

It is totally viable to “divide and conquer” the work for the project, but please consider the following. All great projects have one thing in common regardless of the strategy used to complete the assignment. The group members spent a considerable amount of time making sure that their final presentation has good synergy from start to finish regardless of who did the work and who is presenting. Communication is EXTREMELY important both individually as well as in group settings. It will show up both in the presentation as well in the peer reviews so plan accordingly for an optimal experience with your group members.

Note that many publicly available data set have some data entry errors. While some cleaning is always required, students should not spend all of their time cleaning. The focus should be on addressing the objectives. When it comes to data cleaning, just make sure you communicate what you did, and if there are still concerns with the data you should point them out in the discussions in an appropriate section of the talk. It is totally fine (for this project) to delete variables that are not reliable or have a lot of missing value or delete rows that are suspect. If your team is feeling ambitious, you can play around with imputation strategies (Unit 4). Again, data cleaning should not be the aspect of the project that eats up all of the groups work time.

Confusion matrix

Appendix

Data Set 1: Hospitalization Stays

Objective 1: Hospitals are constantly trying to understand what impacts the duration of a patients stay. The data set consists of 11 variables collected on each of 113 hospitals. One of the variables is the hospitals infection risk, which is an estimated probability of a patient acquiring a hospital related infection. The hospitals are interest to know if the infection risk and related factors are associated with a patient’s length of stay even after accounting for other variables that would/may impact hospital stays such as Age, Number of nurses, number of patients per day, number of beds, routine medical checks, etc. Produce a regression model to address the primary question.

Objective 2: Investigate the data set further with the objective to create a model that can predict as well as possible. In addition to your model from objective 1, fit 2 additional models. One should be an MLR model with added complexity and the other should utilize a nonparametric approach such as KNN, regression tree, or random forest (See Caret help). The group should provide a table with an appropriate error metric to compare the three models considered. Provide a recommendation on which model you would move forward with if the plan was to use this model to predict future patients’ hospital stays.

|  |  |
| --- | --- |
| Variable Name | Description |
| Hospital ID |  |
| Length of Stay | Average length of stay (in days) |
| Age | Average age of patients (years) |
| Infection risk | Average estimated probability of hospital infection |
| Routine culturing ratio | Ratio of # of cultures performed to number of patients without symptoms of infection times 100 |
| Routine chest X-ray ratio | Ratio of # of cultures performed to number of patients without symptoms of pneumonia time 100 |
| Number of beds | Average number of beds |
| Medical School Affiliation | 1=Yes, 2=No |
| Region | Geographic region: 1=NE, 2=NC,3=S, 4=W |
| Average Daily census | Average number of patients in hospital per day |
| Number of nurses | Average number of full time nurses |
| Available facilities | Percent of 35 potential facilities and services that are provided by the hospital |

Data Set 2: MSRP of Automobiles

Objective 1: Understanding relationships between various attributes of a car and their selling price can be very informative for people interested in purchasing a new car. The objective for this analysis is to build a model with relevant variables and interpret the key trends that your model suggests.

Objective 2: Investigate the data set further with the objective to create a model that can predict as well as possible. In addition to your model from objective 1, fit 2 additional models. One should be an MLR model with added complexity and the other should utilize a nonparametric approach such as KNN, regression tree, or random forest (See Caret help). The group should provide a table with an appropriate error metric to compare the three models considered. Provide a recommendation on which model you would move forward with if the plan was to use this model to predict future MSRP values of cars.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| MSRP | Numeric | The response variable |
| Make | Categorical | Make of Vehicle |
| Year | Numeric | Year the car was produced |
| Engine Fuel Type | Factor | Type of fuel the car accepts. Ex: Regular unleaded, Premium unleaded, Diesel |
| Engine HP | Numeric | Horsepower of the car’s engine. |
| Engine Cylinders | Numeric | Number of cylinders in the car’s engine. |
| Transmission Type | Factor | Type of transmission in the car. Usually manual or automatic, but there are a few specialty transmission types in the data. |
| Driven\_Wheels | Numeric | The wheels that are powered by the engine. Ex: Front Wheel, Rear Wheel, Four Wheel Drive |
| Number of Doors | Numeric | The number of doors that the car has. Usually 2 or 4 |
| Market Category | Factor | Various special factors for each car. Ex: Exotic, Luxury, High-Performance, Flex Fuel. Note: we created a new feature using Exotic/Not Exotic for our analysis |
| Vehicle Size | Factor | The size of the vehicle. Ex: Midsize, Large, Compact |
| Vehicle Style | Factor | Body type of the vehicle. Ex: Coupe, Convertible, etc. |
| Highway MPG | Numeric | Fuel efficiency on the highway in MPG |
| City MPG | Numeric | Fuel efficiency in the city in MPG |