Final Project Topic 2: Using Regression

Read through these instructions in their entirety before you begin. These instructions include your research objective, the assignment requirements, and how to report your results. There are several examples of what you should and shouldn't include in your paper. The final section of these instructions consists of a couple of important notes when you submit your work. When you have completed the assignment, you will submit a programming script file and a formal paper document per APA (2020). **You must submit both to earn credit for your work**.

Objective

> Using the data adapted from Huang et al. (2021), considering the vehicle models L200, Q3, CX-5, and XC90, the sport utility vehicle and pickup body types, using either petrol or diesel fuel types, and the six most frequently advertised colors of these vehicles: what are the most influential attributes when predicting the vehicles' advertised price?

Take a minute to consider what you need to address this objective. You need data, of course; what else? Think about the part of that question that provides a *measurement*. The measure-based information in this research question is indicated with the words "most influential attributes when predicting." What analysis methods can be used to gain this information? In lecture 12, which you have not yet seen, you'll be introduced to a method of machine learning that can and will provide that information. You were intentionally allowed to start this assignment before that lecture was available. Most of the time you will spend on meaningful research will be spent cleaning and exploring the data. It is extremely important that this part of your research is always given due diligence.

The data you will use is in the file **`car_ads_fp.RData`** provided in Blackboard (this is not the same data used in lecture 12). When you have finished collecting, cleaning, and exploring the data, the subset you should have when you move to the analysis stage will have seven variables and 3391 observations (rows). Of the seven variables, three are number-type, and four are object-type. If your data does not have these characteristics, it's likely that either you didn't sufficiently clean the data or, somewhere along the way, you did not inspect what you expected.

Requirements

When you are finished, you will submit a formal paper, formatted per APA (2020), and a program script (file type py). You must submit both to earn credit for your work. Your paper will include a **two-to-three**-page narrative. This page count excludes the cover page, reference section, certification of authorship, and appendices. Between the paper and programming, your work will demonstrate your knowledge of every step of the data science research process. The analysis stage of your programming will closely align with the demonstration in lecture 12. The timing of this assignment and the availability of this lecture is intentional.

While you do not yet have access to this lecture, take time to explore the data. Determine the recipe necessary to import the data, create the appropriate subset, and look for data that is not clean. Regarding your paper, without the information in lecture 12, you can write the introduction, objective, sample, and conclusion. Only your results and discussion sections are impacted by the research you do in Python.

When you analyze data in Python, ensure that your Python script file includes your comments regarding the three substeps of this stage of the research process. Additionally, make sure that your programming code demonstrates that you have worked through your analysis plan.

The remaining sections of this document include requirements, information regarding how to document your research, and some useful information.

Reporting the Project

This section of the instructions includes a high-level overview of your paper's required content. You must begin with an **introduction**—a brief summary of the entire paper. You'll need to define the **objective** that this research was driven from, along with the population your **sample** represents (describing what information you used in the analysis, like the vehicle's color or price). You'll need to introduce your method and any weaknesses or limitations of this method in the **method** section. You will use extremely randomized trees regression for this research. (You'll also learn more about that in lecture 12.) To discuss the importance or influence of the vehicle attributes towards your objective, you have to establish the validity of your regression model, along with how your research weaknesses were mitigated. Every result you document must include your interpretations. After presenting the **results**, you need to provide the **discussion**. In the discussion, you'll take the evidence you found and tie that information back to the objective of the research. This is the purpose of the entire project, identifying the answer or answers to the research question and explaining what that means in terms of the information you researched. The last section of your paper is your **conclusion**. Your results and your conclusion are distinct and separate sections. A conclusion does not include any new information. Your conclusion, much like your introduction, is a brief summary of the paper.

What's In; What's Not

This section includes examples of what you should and should not include in your paper and some information about the programming, as well. When you consider what to include in the paper, use the following scenario for perspective.

> Imagine you led a think tank of 20 of the smartest people in the world working on a data science project. It took you and your team 30 12-hour days to finish the analysis. When you reported the results, you were required to summarize the findings and the actions the organization needs to take based on these findings to your manager in under five minutes.

What is the most important information? Do you think that programming code and explaining variables and functions have any business in that five-minute window? (It does not!) It should be difficult to fit everything you need to say in this paper. There won't be room for superfluous or wordy content.

This paper is your delivery. Your paper is not going to contain information like how you felt about the work, the assignment, this class, or your experience while completing this project. The paper will not contain programming code or programming explanations. Lastly, this paper is not going to be written like you're providing someone else directions on how to do the work. When you discuss the information used in your research, use real words. For example, which sentence makes more sense?

> Unacceptable: The sample includes Seat_num four and Seat_num five.
>
> Acceptable: The sample includes advertised vehicles reported as having either four or five-passenger seating.

The unacceptable example is nonsense. When you write your paper or deliver results in any setting, assume the audience has no access to the actual data. Variables used in programming are typically going to sound nonsensical. You must use real words when you discuss the information. Make sure that your paper tells the story about your objective, not programming.

Sample Section

When you describe your sample, tell your audience what it includes, not what you didn't include.

Unacceptable: I deleted a lot of models.

Acceptable: The data analyzed included five vehicle models.

Which one sounds better? Which one conveys more meaningful information? This is a dataset of advertisements. Does it contain every advertisement ever made? How would you document the advertisements you didn't include? You can't! That's just another example of why you describe what you included, not what you didn't. You must share the source of your data, though. When you credit the source of the data used in this research, you'll also need to clarify that this data is *adapted*. (You have not used it in its raw form.) The following is an example of how you may credit the data source you will use. (You can even use this exact sentence.)

The sample data used in this research was adapted from car advertisement data collected by Huang et al. (2021).

As with all citations, you must include the reference in the reference list. (The complete reference for Huang et al. is in the reference section of these instructions.)

Unless you change the population that your sample represents, changes to the data will not be documented in your paper. For example, correcting the assigned data types is a programming requirement. The data and what the data represents are not different when these are changed. Therefore, you will not document that you changed a data type in your paper. If you only used observations representing specific vehicle models or specific vehicle colors that will change the population your data represents, you would report what manufacturers were included in your research. Even though the majority of your time on this entire project will be spent

cleaning data, it's unlikely that anything you accomplish while cleaning the data will be documented in your paper. (In terms of grading, your demonstrated understanding of data cleaning will be reflected in your Python script file's programming code and coding comments.)

Method

The analysis method or methods that led to addressing your research question need to be specified in your paper. Additionally, you must identify any weaknesses or formal statistical assumptions the data must meet. Analyses' weaknesses, weakness mitigation, and formal statistical assumptions are not your ideas. Therefore, when you document this information, you *must* provide evidence (that's going to be an external reference). The reference or references you may use for this part of your work did not do your research, nor did they analyze your data. Ensure you only credit them for *their* ideas. For example, random forest-based models have multiple methods of measuring feature importance. Parr et al. (2018) documented the strengths and weaknesses of feature importance measures (which you will use in this research). Mathew (2022) did research that utilized feature importance to address their objective. This scholar used a fantastic approach to ensure their work was valid and reliable. The scholar didn't specifically write that importance measures had weaknesses. Instead, they described what was done to ensure their measures were valid (see p. 5238, in the first paragraph of section 3.3 in Mathew, 2022).

Results and Discussion

When you write about the results, make sure that you explain why the information you provided matters. For example, you will split the data into training and testing to determine if your machine learning model is valid and reliable. (This mitigates several potential weaknesses.) The objective of the research is not machine learning, nor is it training or testing the model. However, you will need to establish that you did create a model and whether it's valid in order to

meet an objective that requires a model to address. Take a look at the following acceptable example.

> The data were analyzed with extremely randomized trees regression to determine the influence the advertised vehicle's attributes have when predicting the price of a vehicle. Model validity was assessed through training and testing (training: R^2 = .93, *RMSE* = 4324.21; testing: R^2 = .92, *RMSE* = 4824.23). The similar performance between training and testing suggests that the model is both valid and reliable.

This is monotonous content, but very important. The last sentence, the one that answers the 'why,' is the *interpretation* (why as in *why does this matter*). If your interpretation is that it is *not* valid and reliable, you will still report all of the results. You'll need to explain what that means, as well.

In lecture 12, you'll be introduced to all the programming requirements that lead to identifying a characteristic's influence on the outcome. The following information contains an example for the discussion section using information from lecture 12.

> When the model was assessed for permutation-based importance, the year the vehicle was manufactured was the most influential attribute (*permutation scaled* = .76). Only two other features had a large influence, the vehicle manufacturer and the vehicle transmission type (*permutation scaled* = .67 & *permutation scaled* = .38, respectively). The measures of importance suggest that the year manufactured, the name of the manufacturer, and the transmission type are the characteristics that were the most influential in predicting an accurate advertised price.

The last sentence in that example not only identifies why the information matters but also ties this information back to the objective of the research (that's why this is *discussion* content, not

results). This is an extremely important element of any delivery: answering the question, tying your solution to the problem, or connecting the evidence to what it is you are trying to prove.

Programming

In your program, calls to print inspected expectations shall be commented out but not deleted. This is one of the methods used to document your understanding of inspecting and testing your programming code. You may comment out programming code that doesn't work, or if you have coded more than one way to accomplish the same task, you may comment out that code, as well. When your script file contains this type of content, before you submit your work, create a copy of your script file and save it as a revision or version change. Add to the comments at the top of the new script file. Document why the new file was created and what the differences are between it and the original. Using abbreviations, incomplete sentences, and things like slang are perfectly acceptable in your comments, as long as it's understandable. In the revision, delete the programming code you are not using, or that does not work. Again, don't delete the code used to test your code. This type of code should remain in your final version but be commented out. When you submit your work, submit both (or all) versions of your programming script file. Only the final version will be assessed for execution, expectations, aesthetics, and sufficient commenting. The comments at the top of each version should make it extremely clear which version is the final version. If your final version does not meet all of the expectations, but your earlier versions encompassed these missed expectations, you may earn credit for the work you did, despite its missingness in the final version.

Summary

Remember to write succinctly, ensure all you share serves a purpose, tie your research to your objective, and use *real* words. (Your paper will not include programming code,

programming terminology, functions, libraries, modules, or anything else along those lines. There is no reason to even mention Python.) Your delivery is the culmination of the research, so it includes information about your objective and the answers, solutions, or proof necessary to meet that objective. Much of the research work you do between the objective and getting your results is not part of the delivery.

Good to Know

There are a few final elements regarding Blackboard, references, and your submission. When you submit this work in Blackboard, you may receive a warning because the file type of your programming script may not be recognized. Blackboard will still accept your work. If in doubt, use the grade book to validate what files you submitted. The second item to discuss is plagiarism. Whether it's in the narrative of your paper or your programming code, plagiarism is unacceptable. Any external sources you paraphrase or quote must be credited in your paper with citations and references. As stated earlier, you must also credit the source of the data. If you used external sources to assist in your programming, and these sources were not paraphrased or quoted in your paper, they will not be referenced in your paper. If you want to ensure that these external sources are credited, you can refer to them in programming comments in your script file (APA formatting within your script file is not necessary). It’s important to restate that you *must* submit both the programming and the paper to earn credit for any of your work.

References

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

Huang, J., Chen, B., Luo, L., Yue, S., & Ounis, I. (2021). *DVM-CAR: A large-scale automotive dataset for visual marketing research and applications.* arXiv. https://doi.org/10.48550/arXiv.2109.00881

Mathew, T. E. (2022). An optimized extremely randomized tree model for breast cancer classification. *Journal of Theoretical and Applied Information Technology*, *100*(16). https://doi.org/10.14704/nq.2022.20.5.nq22227

Parr, T. Turgutlu, K., Csiszar, C., & Howard, J. (2018, March 26). *Beware default random forest importances*. Explained AI. https://explained.ai/rf-importance/index.html