SDS 322E Project

Part 2 : Classification and Prediction

Overview

You've already explored some data and the second part of the project will ask you to consider relationships with a model to classify/predict an outcome (numeric or categorical) based on some predictors.

For this part of the project, you will fill a quick check in survey on *Canvas* and, once your check-in is approved, you will work on:

- ✓ A narrative report to tell the story of your data analysis (a knitted R Markdown file),
- ✓ A short presentation to share your findings with the rest of the class (recorded video),
- ✓ 3 peer reviews to provide feedback to your peers and learn about different projects!

Again, you might find some inspiration looking at the previous Winning Projects for the USCLAP competition. I highly encourage you to participate to this competition, they have cash prizes!

Working independently or in a group

Stick with your group, join a new group, or keep working individually.
If working in a team, you will be asked to do a few more things (see instructions below) but your team will produce one unique report/presentation. You will also assess each other's contribution to the project in the report.

Topic and datasets

You should continue investigating data from the City of Austin Open Data Portal.

You will have to tidy your dataset(s) as you did for the **Exploratory Data Analysis**. Make sure that you have the following in the resulting/final dataset:

- ✓ at least **50 observations** (i.e., rows)
- ✓ at least **4 variables** but *if a team project, you need at least 6 variables*.
- ✓ at least **1 outcome variable**. Note: some datasets might not have an obvious outcome variable but you can create one (e.g., in the Austin 311 Public Data, we are given the date when the service request was created and when it was closed so we calculate the number of days it took to close the request).
- ✓ at least **1 numeric variable**. Note: some datasets might not have a numeric variable but you can create one (e.g., same example with the Austin 311 Public Data, calculate the number of days it took to close the request).
- ✓ at least **1 categorical variable**. Note: most datasets have at least one categorical variable but you can still create one (e.g., split a numeric value into "small" vs "large" values).

There are no restrictions for the maximum number of datasets, observations, variables, but just be aware that R might run a little slow if you have millions!

Report

The text of your report will provide a narrative structure around your code and outputs with *R Markdown*. Answers without supporting code will not receive credit and outputs without comments will not receive credit either: write full sentences to describe your findings. All code contained in your final project document must work correctly (knit early, knit often)! *If a team project, only submit one report per team.*

Detailed guidelines for the report:

1. **Title and Introduction.** Same as the EDA project and identify clearly what your outcome variable is and what are your potential predictors. What kind of research questions will your data analysis answer? *Make improvements if you had received any suggestions about this section!*

2. **Methods.** Might be the same as the EDA project, unless you need to use your dataset(s) in a slightly different shape which you should document. *Make improvements if you had received any suggestions about this section!*

3. **Exploratory Data Analysis**. Same as the EDA project. *Make improvements if you had received any suggestions about this section!*

4. **Classification and Prediction**. Fit **at least 1 model** (linear regression, logistic regression, k-Nearest Neighbors, or decision tree) to predict **an outcome (numeric or categorical)**, making sure to use a model that makes sense depending on the type of outcome, and based on **at least 2 predictors** in your dataset. *If a team project, each group member fits a different model*.
   - First, fit the model to the entire dataset and then use it to get predictions for all observations.
     - If you are predicting a numeric outcome: calculate the value of the RMSE.
     - If you are predicting a categorical outcome: build a ROC curve and calculate the value of AUC.
   - Second, perform 5-fold cross-validation with the same model. Report the average performance of the model across your *k* folds. *If a team project, each group perform cross-validation for their own model*.

   Discuss the results in a paragraph. How well does your classifier predict new observations? Are there any potential signs of overfitting? *If a team project, compare the performance of the different models*.

5. **Discussion**. Putting it all together, what did you learn from your data?
   - Answer your research question(s) in context, referring to the performance of the model(s).
   - Did the data match what you expected? Anything you are curious about?
   - Consider ethical issues (from data collection to interpretation): what impacts/implications could your results have on the community?
   - If you were going to share your finding with the City of Austin (and you might actually do that!), what would be the main takeaway from your exploratory data analysis? Anything they should know about the state of the dataset (e.g., inconsistency in categories, typos, …)?

6. **Reflection, acknowledgements, and references.** Reflect on the process of conducting this project. What was challenging, what have you learned from the process itself? Include

acknowledgements for any help received: who helped you with the project? Thank TAs, instructors, data owners… *If a team project, that is where you report the contribution of each member: who did what)*. Include references: dataset link(s), a citation for background context).

7. **Formatting**. Create the report using R Markdown, with headers for each section, and comments to the R code. The final report should be no more than 20 pages (the number of pages can vary greatly depending on the cleaning process though). It is extremely important that you **select pages** to corresponding sections when submitting on *Gradescope*. *If a team project, identify all team members on the submission.*

Presentation

You will describe the main steps taken to complete your first project and share some key findings of your exploratory data analysis. *If a team project, only submit one presentation per team.*

Detailed guidelines for the presentation:

- Make slides to present and follow the structure of the report. You do not need to submit your slides. Record yourself while sharing your screen. *If a team project, you can record the video in the same room or through a Zoom meeting for example.*
- During the presentation, introduce yourself, share the title, motivations for choosing this dataset(s), and what are the outcome/predictor variables. Discuss your research question(s) referring to the performance of the model(s). *If a team project, each member shares their own model.*
- The presentation should not last more than 5 minutes. *If a team project, try to split the time equally between each member.* Your video does not have to be high quality, but we should be able to hear you and see your presentation materials clearly throughout the entire video.

Detailed guidelines for the peer reviews:

- Each student will peer review 3 videos from other classmates/teams (it also means that 3 other students will watch your video). Peer reviews will be assigned the day after the presentation is due and will be due within a week the presentation was due.
- You will use the rubric (see below), evaluating the Materials, Presence, Quality for the video presentation. Write individual and constructive feedback:
  - ✓ Be specific in your comments: refer to a specific phrase, visualization, or slide.
  - ✓ Offer a solution to improve the work: mention possible revisions or links to helpful resources or examples.
  - ✓ Be kind: all your comments should be framed in a supportive way.

  Consider the following:
  - × If you leave the same comment to all 3 of your peer reviews, you won't receive full credit.
  - × If you assign a grade but don't leave any comments, you'll receive some credit.
  - × If you leave comments but don't assign a grade, you'll receive half credit.

Grading Rubric

Your project will be assessed as follows:

Rubric Item	Description	Points
Check In	Decide if you are completing the project independently or with a team. Share your choice of dataset(s), the outcome and predictor variables.	**10**
Report	**Title/Introduction**: present the context of the dataset(s) and share the research question(s) your exploratory data analysis will aim to answer.	**5**
	Methods: document the process to get the dataset in shape for data analysis.	**5**
	EDA: create visualizations and summary statistics to investigate your dataset.	**5**
	Classification and Prediction: create a model to predict your outcome and check the performance of your model using cross-validation.	**60**
	Discussion: answer your research question(s) in light of your results and ethical considerations.	**10**
	Reflection, acknowledgements, and references: reflect on what you have learned and refer to any external resources.	**10**
	Formatting: well organized and easy to navigate.	**5**
Presentation	**Materials**: All required elements were included. Texts and figures are easy to read, free of spelling errors. The information is presented in a logical structure.	**10**
	Presence: Presenter(s) is audible, introduces themselves, shows enthusiasm for the topic and seems prepared. *If a team project, the time is (approximately) split equally among the team members.*	**10**
	Quality: At the end of the presentation, the audience should have a good understanding of how the results can be interpreted in context. The visualizations and statistics were appropriate and conclusions are not overreaching.	**10**
	Complete 3 peer reviews using the rubric and leaving constructive feedback.	**10**
	Total	**150**