Stat 172 Final Project (Group)

Dr. Lendie Follett

2025-11-05

You (as a group of 2-3) are tasked with finding a sufficiently complex labeled[1] data set. **As before, I must sign off on this data set to ensure appropriateness.** You must invent a 'client' for whom you are doing this analysis. As is often the case, your client is interested in both understanding the relationships in the data and being able to predict future observations.

Your client requires:

1. A prediction tool that can be used to predict a 2-level (equivalently, binary) categorical outcome on *new* data. Your client is fairly savvy, so you'll have to convince them that you've thoroughly evaluated it.[2] (predictive model)
2. A parsimonious[3] GLM that can be used to help understand important relationships between the explanatory variables and your Y variable. [4] (descriptive model)
3. Stupendous, meaningful data visualizations that strengthen findings of predictive and descriptive models. (to aid in describing relationships)

Thus, you will build both a descriptive (e.g., GLM) and predictive (e.g., random forest) model (not necessarily in that order) and present the results as if you were presenting to your (again: tech savvy) client. Your goal is for your final models to be put into automated production to aid in both forecasting and understanding. [5]

The final deliverable will be

- a 10-15 minute presentation describing your methods, findings, and advice. *12 minutes is NOT a lot of time. Be concise. Know what you need to say and say it. See last page for recommendations on presentation format.*
- README file sufficient to reproduce your results
- R[6] code files sufficient to reproduce all of your results given the raw data. (you will likely, I would hope, have multiple R files. One for cleaning, one for descriptive, one for predictive, etc... your README should walk a reader through your process.)

[1] See definition in notes.

[2] Transparently and fully describe your best tool's performance on test data. For some data, creating a 'good' prediction tool is impossible, but what I really care about it whether you're able to tune it and validate it correctly and describe this honestly to another human.

[3] What do I mean by this? Thoughtful and objective x variable selection. We know several ways of how to do this *well*.

[4] In real life, it is common to have a descriptive model to supplement the predictive one in order to gain insight.

[5] Note: Excel is never part of an automated process. e.g., data cleaning in Excel is not ok.

[6] and SAS, if you chose to fit any GLMs in SAS instead of R

RUBRIC

Visualization

Outstanding (20)
Modern and quality statistical graphics using modern R packages are used to enhance analysis. Thoughtful color palettes, labels, titles, and plot types are chosen. Multivariate charts (multiple x's along with y) are incorporated if appropriate. This helps the audience better understand the data and the problem at hand and your insights. Visualization is ultimately tied to methods chosen; variables involved are chosen for a good reason.

Acceptable (10)
Visualizations are modern and generally high-quality but lack at least one component of quality (default colors, default labels, inappropriate plot types, etc.) and/or seem unconnected to analysis.

Needs work (0)
Visualizations are either not done or are poorly integrated. If done, visualizations seem an afterthought and do not enhance analysis.

Predictive Methods

Outstanding (25)
Predictive models are correctly chosen, tuned, interpreted, and implemented. Multiple predictive models are compared appropriately. Out-of-sample (testing) metrics are appropriate and meaningfully described in the context of the data and research question. Code automation is fully incorporated.

Acceptable (15)
Models are correctly tuned. However, tuning procedures are reported but may lack documentation/clarity. Some out-of-sample metrics may be missing or are interpreted incorrectly and/or lacking meaning to a broader audience. Or, only one predictive model is attempted.

Needs work (0)
Model fitting/tuning is either not done, done incorrectly, or not automated.

Descriptive Models

Outstanding (20)
Random component AND systematic component have been chosen thoughtfully. In particular, variable selection is done using appropriate methods. Further, interpretations of coefficients corresponding to both numeric and categorical variables are technically correct and meaningful. Uncertainty around interpretations is quantified and presented appropriately. Interpretations are tied back to the original goals/interests of the client. Predictions are put into context through actionable advice.

Developing (10) Material presented is technically accurate but interpretations are not tied back to the original goals/interests of the client[7]. OR, there is some aspect of the model that is inappropriate.

Needs work (0)
Interpretation of coefficients is incorrect or model is flawed to the extent that it cannot be used to provide insight.

Data Mining Enhancements

Outstanding (15)
Project makes use of additional data mining/reproducibility techniques in a way that is appropriate, relevant to the problem, and creative. This might include at least one of:

[7] for example, you have a slide titled "coefficient interpretations" and your client is left wondering what this has to do with them and why they are here.

- Novel graphics (e.g., a choropleth map? a hammock chart?)
- Use of clustering
- Text mining (*You can scrape any subreddit you want! I can supply the code for this.*)
- RMarkdown slides (correctly - ask me what I mean by this before you go this route)
- Correct use of Github with a helpful README file (again - ask me what I mean by this before you go this route)
- Interesting use of scraping web data

Acceptable (5)
The student makes use of additional data mining techniques, but it seems disconnected from the rest of the analysis and/or is somewhat inappropriate in the data context and/or lacks effort. For example, "For our Data Mining Enhancement portion of the rubric, we did clustering. Here are our clusters. Moving on."

Needs work (2)
Incorrect use or missing.

Presentation

Outstanding (10)
Information is verbally presented clearly and professionally. Presentation materials (PowerPoint or otherwise) are easy to read and professional. Most importantly: *presenters are mindful that they are speaking with a client and align their material and dialog accordingly.*[8]

Needs work (0)
Verbal communication is unclear or seemingly unpracticed. Slides/materials are difficult to read in video.

README file

Your README file should be either a plain text (.txt) file or a markdown (.md)[9]. It should have the sections/features/qualities outlined in the rubric below.

There are many examples of README files on your favorite Github repositories. Here's one to give you an idea of (1) what a .md file is and (2) what you might want to put in it: https://github.com/LendieFollett/Hybrid-Targeting.

Outstanding (10) Project Overview: Clearly describes project goals, methods, and data sources in a succinct summary. Conveys purpose and utility of the project, making it easy for others to understand at a high level. **Data Sources and Preparation**: Provides comprehensive details on data sources, format, and preparation steps. If data is external, offers clear access instructions. Data processing is done via code. Any non-code pre-processing is detailed. Lists all needed packages, with instructions on setting up the environment. Instructions are complete and error-free, enabling seamless setup. **Code execution**: Gives clear, step-by-step instructions on how to execute the code. Lists required commands and settings, with examples. Code is ready for reproduction by others. **Organization**: README is well-organized, free from grammar and spelling errors, and professional in tone. Uses technical language appropriately for a data science audience.

Acceptable (5) At least one of the above aspects of the README is missing or seriously flawed.

Needs work (0 pts) Multiple aspects of a README missing to the point that the README is not useful to reproduce the group's work.

Note on Generative AI: You can use it, but I want to know exactly where and how you used it. In other words, cite your use of it.

[8] Here's an extreme example of what I don't want: "Here is our data mining enhancement… Here are our coefficient interpretations… Here are our plots…" No, that's obviously inappropriate and everyone is uncomfortable.

[9] note that this is different than a RMarkdown file - a .md file will automatically be rendered in Github while a .Rmd will not.

Structuring your presentation

What I don't want? 1. Data 2. Model 3. Coefficient interpretations 4. ROC curve 5. Plots 6. Thanks for listening! :)

That's unprofessional, largely because it's *centered around the data analyst.* Instead, choose to center your analysis around your client and their business problem.

Here is a possible way to structure your presentation. You don't *have* to follow this order… but you should have a good reason if you choose not to. You'll see in some places where this links back to the rubric.

1. **Title & Context (1–2 minutes)**

Project title and student/organization names.

Context or motivation: What business question, problem, or opportunity led to this project?

Example: "Rising customer churn in the subscription service has created uncertainty in forecasting revenue."

Goal statement: "Our objective was to understand key drivers of churn and develop a predictive model to identify at-risk customers."

Tip: This is your time to convince people that they care about your presentation. Or lose them entirely.

2. **Business Problem** 1 minute

Define the problem in business terms, not statistical jargon. What decisions will this analysis inform? What are the key questions being addressed (2–4 bullets)? e.g., Is it possible to preemptively predict customer churn? Which customer-company characteristics predict churn? Which of these are actionable and how should we act? How can we allocate retention resources more efficiently?

Tip: Clarify early how success is defined — e.g., "A good outcome is a model that identifies 80% of churners before they actually churn." Important Note: It's ok if you don't meet your own threshold. You're not being graded on your actual sensitivity/specificity. You're being graded on how you built your model and how you measure and explain your model performance. Some stuff just isn't predictable. See rubric section **Predictive Models**.

3. **Data Overview**. 1 minute

Source(s): Where did the data come from?

Basic structure: Number of observations, relevant variables, and time period.

Data quality / preprocessing steps (only the essentials: missing data handling, aggregation, etc.).

Tip: Keep this concise; show a quick (but quality) visual or simple table to build trust in your data processing.

4. **Methods** 3-4 minutes, likely several slides

Explain how you approached the problem in plain language:

"We compared several predictive methods including a random forest, and a logistic regression fit with MLE, lasso, and ridge penalization."

"We used hierarchical clustering to segment customers based on demographics."

Describe evaluation metrics: accuracy (unlikely, I would hope), sensitivity, AUC, etc.

Emphasize *why* you chose the method — not just what you used.

"We chose logistic regression because interpretability was a priority. We supplement our descriptive insights with a random forest because that had the best predictive power." (recall, I'm asking you to give me both a predictive and a descriptive model…)

Tip: Focus on intuition and relevance to the business goal. Don't let the client forget why they (specifically) are listening to your presentation.

See **Predictive Models** and **Descriptive Models Data Mining Enhancements**.

5. **Results** 3-4 minutes

Tell them: Start with high-level findings first ("What did we learn?“), then support with evidence.

Use clear visuals: meaningful plots with x and y, variable importance (qualituy), predicted vs. true on test, etc. See **Visualization** section of rubric.

Highlight 3–5 key insights. Example: "Discount is the strongest predictors of churn - but, based on the plot, only up to 20% discount matters."

Tell them again: Translate each insight to business meaning or actionable takeaway.

"By targeting high-tenure customers with 15-20% discount code mailers, we could reduce odds of churn by 15%." (sounds like an odds ratio to me)

6. **Recommendations** (*tell them a third time!*) 1-2 minutes

"We recommend. . . ."

Include a visual roadmap or priority list if relevant.

Tip: Always connect back to the original problem and goal — close the loop. This is NOT about you and it is NOT about your analysis - it is about your client.

7. **Limitations** 1-2 minutes

Acknowledge limitations transparently - this is the difference between a good presentation and a great one:

Data limitations (sample bias, missing features).

Methodological limits (correlation isn't the same causation).

Poorer than expected model performance. For example, "Given 100 *positive events*, we only expect to correctly identify 64 of them before they occur based on our model."

Explain the heck out of your sensitivity and specificity - in % terms, in count terms, in practical terms, etc.. **Predictive Model** - this section in the rubric doesn't just condern your code, it also concerns how you talk about your model performance.

8. **Conclusion and questions** 1 minute

Summarize in one slide. Don't ramble, though it can be tempting.

Problem (1 sentence) then Method (1 sentence) then Key insight then Action.

End with a takeaway statement. "Our model showed that we can use customer data to predict a future churn event; though, only up to a point. Still, we recommend blah blah blah."

Questions?