PACE UNIVERSITY

CS 667 Practical Data Science

Presentation & Meeting Skills; Model Selection

Spring 2025

Stephanie Langeland, M.A.
Adjunct Assistant Professor
Seidenberg School Of Computer Science and Information Systems, Pace University

Attendance

Agenda

- Presentation and meeting skills
- Model selection

Elements of good presentations and meetings

- Presentations and meetings should be conversations, not lectures
- Know your audience so that you can engage with your audience
  - Ask if there are questions
  - Look around to see how people are reacting
  - Make eye contact
  - Smile
  - Tone and pitch of voice

Elements of good presentations and meetings

Meeting materials

- Concise
- Clear
- Correct
- **Review to the example template in the appendix - you will use this template as your template for decks you make for this class**

Running successful meetings and presentations

- Always include an agenda in every meeting invitation
  - People are busy, they need to know which conflicting meetings to accept
- At the start of every meeting:
  - Thank everyone for their time
  - Set expectations
    - Explain the agenda
    - Explain the scope
- Stay within the scope of the agenda
- Be on time starting and ending the meeting
  - People with back to back meetings cannot stay late
- When someone interrupts you:
  - Be cordial
  - Welcome questions and comments
  - Stay on track
- When someone asks confusing questions or makes confusing comments:
  - Ask for clarification ➜ understand what they're asking and why, so that you know whether this should be addressed now or offline in a separate conversation

Math speak

Class exercise: math speak

Scenario	Option A	Option B
Random forest model that classifies dog breeds: fur color ranked #3 in feature importance	An advanced decision tree method concluded that fur color ranked #3 in feature importance	When predicting a dog's breed, we found that fur color is an important variable

Class exercise: math speak

Scenario	Option A	Option B
Explain training data vs. validation (hold-out) data vs. prediction data. You can use the dog breed classification model as an example.	Training data is a large dataset used to teach a machine learning model how to make predictions or perform a task. Validation (hold-out) data is a portion of historical, labeled data that is intentionally excluded from the data sets used for training and validating supervised machine learning models. Prediction data is a dataset that is missing the outcome variable's actual values. A model applies its learnings from training to this dataset to make predictions.	Next slide

Predictor Variables			Outcome Variable		
Fur Color	**Temperament**	**Size (Male)**	**Breed**	**Dataset**	**Procedure**
Black	Gentle	65-70 pounds	**Greyhound**	Training	Model learns on these data.
Blonde	Energetic	65-75 pounds	**Golden Retriever**		
Tan	Powerful	35-60 pounds	**Pit Bull**	Validation (hold-out)	Model applies learnings from training and predicts dog breed. We compare predictions vs. actuals to understand training accuracy.
Brown, white, and tan	Affectionate	20-30 pounds	**Beagle**		
Sable and white	Independent	45-60 pounds	**Unknown - model needs to predict breed**	Prediction	Model applies learnings from training to predict dog breed when it is unknown.
Brown and white	Charming	140-180 pounds			

Class exercise: math speak

Scenario	Option A	Option B
Explain class imbalance. You can use the dog breed classification model example.	Class imbalance means that there is not an even ratio of classes in the training data.	Class imbalance is when we have more cases of Golden Retrievers than Labs in the training data.

Class exercise: math speak

Scenario	Option A	Option B
Explain accuracy vs. f1 score	When we measure how well or poorly a model performed, we calculate the model's predictive accuracy. There are 2 main metrics we rely on: 1) Accuracy, which is purely how many predictions the model got right. 2) F1 score, also measures how many predictions the model got right but it accounts for class imbalance. We use f1 score instead of accuracy when there's a class imbalance in the training data.	The accuracy evaluation metric measures how many correct predictions a model makes compared to total predictions. Whereas the f1 score conveys the balance between the precision and recall.

Resources

- 25 Tips to Instantly Improve Your Data Visualization Design
- 8 Tips for Creating a Compelling Presentation for Data Science
- How to Present a Data Science Project
- 7 Tips for Delivering a Great Data Science Presentation

10 minute break

How do you choose the appropriate type of model for a task?

Investigate before you choose a specific model

- Ask for context on the task at hand - why do we need to build this and who is the end user
- What's the deadline? Is this a long-term project or fire drill to be done right now
- What's the final deliverable? Work backwards to plan out:
  - Model type
  - Data
    - Availability
    - Sample period
  - Tech stack / computing power
  - Know your end-user and what they're comfortable with in terms of ML

Choose the appropriate model for the task

Scenario	Choose the model
You're on the supply chain analytics team at a large consumer products company. One of your largest retailers has reported several instances of stockout or overstock situations. ● Every time there is a stockout, your company loses potential income from customers who would have bought the product. ● Every time there is an overstock, the company misses an opportunity to efficiently re-distribute products to other local retailers. Your task is to build a model that will predict future inventory levels. Using those predictions, alerts can be created to warn you about a future overstock or stockout situations. These alerts will allow the warehouse team to re-distribute products efficiently to stores ahead of the predicted overstock or stockout.	● K-means ● ARIMA or seasonal ARIMA ● Linear regression

Choose the appropriate model for the task

Scenario	Choose the model
You work for a property & casualty insurance company. The Actuaries who are insuring restaurants note that they often find helpful insurance risk information in the Yelp reviews of restaurants. E.g., reviews about slip and falls or grease fires help them understand unreported insurance risks. They often factor the findings of this research into their premium calculations. Your task is to help automate this research process: ● Create an automated way to search through each restaurant's Yelp reviews. ● Tag each review that mentions an insurance risk and decide which risk is predominant. ● For each restaurant and insurance risk category, calculate a risk score based on structurally negative reviews.	● Lasso regression ● NLP similarity and sentiment models ● NLP entity extraction and sentiment models

Choose the appropriate model for the task

Scenario	Choose the model
You work in HR analytics. Your team sent out an anonymous employee experience survey. The survey questions asked a variety of questions designed to understand: ● Rate your happiness at work. ● Rate your level of engagement at work. ● Rate how much access you feel you have for opportunities for advancement within the firm. ● Overall: Rate how likely you are to stay with the firm for the next 2 years. All responses were recorded using a Likert scale. E.g., Rate your happiness at work: 1 = very unhappy; 2 = unhappy; 3 = neutral; 4 = happy; 5 = very happy. Your task is to split up the results into different groups: by race, gender, and tenure. For each group, predict the likelihood of turnover, i.e., that someone will leave the firm in the next 2 years.	● Logistic regression ● K nearest neighbors ● ARIMA

Appendix

NOTE FOR STUDENTS: Use this deck as your template for your Data & EDA, Methods, Findings, & Recommendations, and Final presentation decks.
This template aligns with the rubric provided for those assignments.

Project name: phase of project

Date of Presentation
Student Name
Student's Pace Email Address
Class Name: Practical Data Science
Program Name: MS in Data Science
Seidenberg School of Computer Science and Information Systems
Pace university

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modeling methods
- Findings
- Business recommendations and technical next steps

Executive summary

Summarize the problem in 1-2 sentences.

Summarize the solution in a few sentences or bullet points.

Project plan recap

Deliverable	Due Date	Status
Data & EDA	MM/DD/YY	Not started / In Progress / Complete
Methods, Findings, and Recommendations	MM/DD/YY	Not started / In Progress / Complete
Final Presentation	MM/DD/YY	Not started / In Progress / Complete

Data

Data

What goes on this slide:

- Data details, for example but not limited to:
  - Data source
  - Sample size
  - Time period
  - Data that was purposefully included or excluded
  - Important notes about the data that need clarification
  - If/when applicable: provide a hyperlink any appendix slides that provide supplemental information for each slide
- Assumptions
  - E.g., if you're looking at dog licensing data, you'll assume that most dogs are licensed and therefore captured in the dataset because your external research showed that dog licensing is enforced for the locations covered by the dataset

For final class presentation:

- This section should be no more than 2 slides (1 slide is acceptable)

Exploratory Data Analysis

Clear and concise title that gets to the “so what” of the slide

What goes on this slide:

- Data visualization and short summary of what it means (AKA the “so what”)
- If/when applicable: provide a hyperlink to any appendix slide(s) that provide supplemental information for each slide

For final class presentation:

- This section should have 5 data visualizations: 1 data viz per slide that follows the above formats
  - The data vizes should relate to your model’s outcome variable and features as this deck should tell 1 cohesive story
  - If you want additional data viz slides (although not required), you can put them in the appendix
- You do not need to have a different type of data viz for each slide. The purpose of each data viz is to portray the “so what”, not to show us that you know how to make different types of figures. So choose the type of data viz that appropriately portrays the “so what” clearly. Overly complicated data vizes do more harm than good - most of the time, simple charts get the job done, e.g., bar charts, scatter plots, line charts, etc. Refer to the data viz slides from the “Data Science Project Setup; EDA” class for guidance.
- This section should be no more than 5 slides

Modeling Methods

Modeling methods

What goes on this slide:
Details about the model you built, for example but not limited to:

- Outcome variable - this is the “so what”/key takeaways ➜ this is what you’re trying to predict, so the story you’re telling needs to always come back to explaining your outcome variable
- Features - explain the rationale/hypothesis behind choosing each feature
    - If it’s a long list of features, you can summarize them into more general groups, e.g., “features related to consumer behavior (refer to this slide in the appendix for the full list)”
    - If it’s a short list, then you can list them here
- Model type and rationale for choosing this type of model: create 2 versions of this slide:
    - Non-technical version goes in the main deck in this section: For a non-technical audience, this main slide should explain the model using examples and in layman's terms. Do not do not explain the model in a theoretical way. Additionally, you should hyperlink to an appendix slide (see next point below) with additional technical details.
    - Technical version (PUT THIS VERSION IN THE APPENDIX): For a Data Scientist audience, this slide should go into great detail about the model on a theoretical and applied level.

Important note: know your audience - these types of slides need to be written differently for technical vs non-technical audiences.

Reminder: At this point, you are preparing the stakeholders for the model you are GOING to build. So at this point, you should not be talking about model results yet. In the next section, you’ll talk about model results/findings.

For final class presentation:

- Organize this section’s topics in the order provided above
- This section can be up to 3 slides
- You will only build 1 model for this assignment. Therefore, you will only explain that type of model.
- Remember, you’re presenting to a non-technical audience of business stakeholders

Findings

Title that conveys the main point

What goes on this slide:

Details about the modeling findings, for example but not limited to:

- Explain your results with visuals and in clear English - connect **every** result back to the business problem
  - “Non-findings” are also findings
- Example of expected results readout for a linear regression model, but not limited to:
  - Adj R-squared and what it means.
  - Add a plot of actuals (points) vs. line of best fit and explain it.
  - Coefficient explanations: With each <1-unit> increase in <X1>, there is an <increase/decrease> of <coefficient value> in <Y>, holding all else constant. <Feature name> <is/is not> statistically significant, which means <there may/may not> be a relationship between <X1> and <Y>. Add comment connecting this finding back to the business problem.
    - Pro tip: A bar chart with the coefficient magnitudes help support these points (X-axis = unit, Y-axis = feature name, bar color = color A means coeff is stat sig vs. color B means that coeff is not stat sig).
    - Can add hyperlink to an appendix slide that has the full ANOVA table and other important metrics, for the Data Scientists in the audience.
  - Add a comment about your final verdict on this model - would you recommend putting this model into production to solve your business problem or not? Explain why. This gets your audience ready to hear about the data science next steps, which you’ll discuss in the next section.

For final class presentation:

- You can decide how many slides to make this section - remember to practice your presentation so that you don’t go over the presentation time limit

Business Recommendations & Technical Next Steps

Business Recommendations and Data Science Next Steps

What goes on this slide / for final class presentation:

Recommendations

- Choose your top 2 most important model results findings (the most important findings directly connect back to the business problem and have actionable insights)
    - For each important finding: connect it back to the business problem and give at least 1 actionable recommendation that the business can execute

Technical next steps for the Data Science team

- Make up next steps for this fake project. E.g.:
    - Should you build a more advanced model to answer open questions that couldn't be answered by this project?
    - Should you collect more data to expand the scope of this project to address a different business problem that you know is connected / relevant?
    - Are you recommending this model for production?

This section can be up to 2 slides.

Appendix

Project Materials

- Git Repo: <link>

Title that connects back to the main slide in the deck

- Details
- Note: make sure that the main slide has a link to this slide because you don’t want your audience wondering what main slide this appendix slide supports