Project Proposals

Project Pitches & Evaluations

Key dates:

project proposals due - 9/26

peer project evaluation due - 10/3

staff release approved projects - 10/8

In this first phase, you will **propose a project** that aligns with your personal, professional, and academic interests and passions. Once the proposal deadline has passed, you will then **evaluate the proposals** made by your peers.

Allowing you to propose your own projects, rather than selecting from a predefined list, will enhance your engagement and lead to better learning outcomes. This approach will also foster your independence, critical thinking skills, and creativity, preparing you for real-world scenarios where you may be required to initiate and lead your own projects. Call on your inner data scientist and take charge of your project experience! Many groups in the past have continued their work on their project post our class, and taken their papers to conferences. Think of this opportunity as your starting point!

The project proposal and peer evaluation **do not count towards your final grade,** but it is a crucial first step for the beginning on any project. All final projects will be based on the project proposals by this deadline, so failing to submit a proposal will restrict you to working with a topic that was proposed by other students. For this reason, **we highly encourage you to submit a proposal so that you can work on a topic you are interested in.**

Grouping

Projects can be *proposed* by individuals or by a group (see 'deliverables' below). But the projects themselves must be tackled in groups, and students will be required to join or form a group of **3 to 5** students by **10/17 (MS1)**. If there are any issues in finding or forming a group, the teaching staff can help resolve them. There will be **no** individual projects.

Proposal

Project proposals should be concise and to the point. We are not expecting any group to produce pages and pages of well-written details! Rather, at its core, the proposal should be a unique and innovative idea that presents a data science problem in an engaging and interesting manner. The proposal should make use of any of the methods covered over the span of the course. Finally, ensure that the scope of your problem/question is clear - there should not be any doubts as to what problem/question you are proposing to solve.
In summary, the proposal should include the following information:

Title and Authors:

A good title goes a long way in attracting your audience's attention. A creative title that is informative and relevant is important. Include the title of the project, your names, and your email address

Background and Motivation:

Briefly describe your reasons for choosing the topic of interest, including any prior background, research interests, or reading (papers, blog posts, etc.) that prompted you to propose your topic.

Data:

Data are an essential part of any data science project, and hence finding reasonable data to work with will be paramount. As part of the proposal, you must provide a source(s) for the data that you will explore in the project, including the data source, a description of the data set, and key attributes of the data set, and its relevance to the problem you which to explore.

You are not expected to engage in data cleaning at this stage; however, do indicate if there are any

foreseeable problems with data quality that may require substantial work in the exploratory data analysis phase (missingness, merging of datasets, etc.).

*Project proposal without references to **publicly available**, relevant data will not be accepted.*

Scope:

With regard to the scope of your work, we leave this largely to you. The project can be as simple or complex as you want it to be, regardless of the topic you choose. However, note that the goal of the project is to demonstrate thoughtfulness in how you approach an interesting problem. For instance, do not use a model/algorithm nobody in the group understands simply because it performs better. Rather, the project should be a place to practice what you learned, build upon lecture ideas, develop research skills, and have fun!

Groups with students enrolled in AC 209a will be expected to utilize some method not explicitly covered in class in their project. This will require some outside reading. These groups should be able to communicate an understanding of this new method as well as its relevance to their project in their final report at the end of the semester.

There is no special AC 209a requirement for the project proposals themselves.

Deliverables:

Submit a PDF or word document of your proposal on Canvas.

If you have a tentative group with whom you plan to work on the project, have all group members join a pre-made "Project Proposal" group on Canvas and select one member to submit the proposal on behalf of the group. Go to **Canvas --> People --> Project Proposal Groups** and have members join a group. One member who submits in this group will submit for all group members.

These groups are not set in stone. Students will be free to change groups and project choices from among the final approved projects up through **10/19.**

Peer Evaluation

YOU will evaluate projects proposed by your peers.

Assignment: You will have a maximum of 4 projects that your peers proposed, and you will evaluate them.

Expectations: An objective assessment and ranking of the projects. You will rank the projects based on your preference, proposal clarity and data quality and data availability (provide brief comments if necessary). You can choose how you may want to evaluate the proposals as a team, but we suggest that you discuss the projects and vote on them (it should not take more than an hour).

How: You will be assigned proposals to evaluate through the same Canvas assignment for the proposals.

Staff Review

After the peer project evaluations are in, staff will review the proposals and release the final list on **10/10**.

Example Proposals

(note: actual proposals should include links to their proposed datasets)

Example Project Proposal 1

Project Proposal Title: Predicting Legal Case Outcomes Using Decision Trees and Random Forests

Group Members:

Pavlos Protopapas (pavlos@email.com)

Kevin Rader (kevin@email.com)

Chris Gumb (chris@email.com)

Background and Motivation:

In the legal field, predicting the outcomes of cases based on historical data can be valuable for legal professionals. While legal scholars and lawyers often rely on case precedents, machine learning algorithms can aid in making outcome predictions based on patterns in case data. Our goal is to use decision trees and random forests to build models that predict the outcomes of legal cases based on various features, such as the type of case, legal arguments, and court jurisdictions.

Data:

We will use a dataset from the Cornell Legal Information Institute (LII) containing details of U.S. Supreme Court decisions. The dataset includes features such as case type, legal arguments, and rulings. We will also explore merging this dataset with other publicly available datasets on lower court decisions. The data is structured, but we anticipate potential issues with missing values and merging datasets from different sources.

Scope:

We will begin by exploring decision trees to identify key factors that influence case outcomes. After tuning the model, we will move on to random forests for improved accuracy. This project will allow us to practice supervised learning techniques and gain insights into patterns in legal decisions that may assist legal professionals.

Example Project Proposal 2

Project Proposal Title: Identifying Factors Influencing Housing Prices Using Multiple Linear Regression and PCA

Group Members:

Ivy League (ivyleague@email.com)

Ella Vator (ellavator@email.com)

Sam Sung (samsung@email.com)

Background and Motivation:

Housing prices are influenced by a variety of factors, from location to home features, economic conditions, and more. Understanding these factors can help homeowners, real estate professionals, and policymakers make better decisions. Our goal is to use multiple linear regression to identify the key factors driving housing prices. Additionally, we will employ principal component analysis (PCA) to reduce the dimensionality of our dataset and focus on the most significant variables.

Data:

We plan to use publicly available housing data from the Zillow Research Data, which contains information on housing prices, property characteristics, and market trends across different U.S. regions. The dataset includes details like home size, year built, and proximity to amenities. We anticipate needing to handle some missing values and potentially derive new features to improve the analysis.

Scope:

We will start by performing an exploratory data analysis to identify important features. Next, we will use multiple linear regression to quantify the impact of each feature on housing prices. PCA will be employed to reduce the feature space, helping us understand which features are the most critical for price prediction. This project will enable us to practice multiple linear regression and dimensionality reduction techniques on real-world data.