Report

Submit Assignment

Due Jun 3 by 11:59pm **Points** 55 **Submitting** a file upload **File Types** doc and pdf

1. Introduction

Please choose one of the projects below for your final project. You will be responsible for handing in a report of your project. This report should not be any longer than 4-5 pages (though if you have a lot of graphs it might be longer). Please see sections below for more information.

Note on plagiarism:

The datasets we will be using are public datasets. As such, they have been used for research and some have had papers published about them.

Feel free to use papers and outside research to help you guide your work. However, it is not acceptable to pass off results as your own if you did not replicate the work, or to use text/tables/etc from other sources without crediting that work.

Example of *non*-acceptable use of online/published resources:

Presenting results from another paper/resource as if you did the work yourself.

Example of acceptable use of online/published resources:

Replicating techniques done in a published paper/online resource, citing that resources, and reporting and comparing the results of your implementation to those of the original resource.

The goal of the project is to get practice at analyzing datasets, not to come up with new, novel techniques. As long as you learn something new, report what you learned and cited any resources you used, you will be fine.

2. Report Format

The paper should be around 4-5 pages and have the following sections.

Note: Although you need to compare at least two methods, you may choose to compare more than two for bonus points. These bonus points will be applied to your midterm.

1. Authors (group members) [1 point]

2. Introduction: Describe your problem and the data [6 points]

3. Method A: Describe the first method you tried and why (e.g., linear regression) [10 points]

4. Method B: Describe the second method you tried and why (e.g., random forest of regression trees) [10 points]

5. [Optional] Method C: Describe a third method [10 bonus points]

6. Results: How well did Method A and B work? Use some standard form of comparison, e.g., MSE, accuracy, etc. If you had to choose between A and B, which is better and why? [Optional: If you have method C or more, you must also compare them to A and B to obtain bonus points]. [20 points] **NOTE:** when comparing results, make sure to compare both accuracy/model fit and how *interpretable (i.e., what features are important)* the model is.

7. Conclusions and Learnings: Summarize what you learned. [5 points]

Note: Your writing will be evaluated too (Is the writing clear and easy to read? Is the English usage professional?) [3 points]

Some examples for Methods A and B:

Method A: Logistic Regression / Method B: Random Forests

Method A: Mutlivariable Regression / Method B: Lasso / Method C: Ridge Regression

Note: If you have taken other related classes, feel free to use other methods that were not discussed in class.

3. Projects

Please choose one of the following datasets and projects.

Project 1. King Country Housing Prices

https://www.kaggle.com/harlfoxem/housesalesprediction (https://www.kaggle.com/harlfoxem/housesalesprediction)

This Kaggle dataset contains house prices for homes sold in King County along with a number of features. You should design a model to predict the housing prices. If you need some help getting started, it is a good idea to start with a regression model. This might be good enough, but you should evaluate the fit and determine if this is an appropriate model.

The goal here is two-fold:

1. To predict the house sale prices given the data set.
2. To learn what features are the most important in the sale price of the house (i.e., If I wanted to invest in a home, what features are most important to maximizing its resale value?).

Project 2. Forest Cover Prediction

https://www.kaggle.com/c/forest-cover-type-prediction/data (https://www.kaggle.com/c/forest-cover-type-prediction/data)

This Kaggle Dataset contains information on forest cover. In this project we use classification to determine if the forest cover is one of the following based on a number of features. To get started we might first choose logistic regression and evaluate the fit by testing the accuracy on some hold out data.

1 - Spruce/Fir
2 - Lodgepole Pine
3 - Ponderosa Pine
4 - Cottonwood/Willow
5 - Aspen
6 - Douglas-fir
7 - Krummholz

The goal is two-fold:

1. Predict the type of forest cover given the features in the data.

2. To learn what features are important in the classification of forest cover (i.e., what features of the landscape seem to be make it most preferable to the various species of trees).

Project 3. Random Acts of Pizza - Challenge Problem

https://www.kaggle.com/c/random-acts-of-pizza/data (https://www.kaggle.com/c/random-acts-of-pizza/data)

Note: As several features are text fields, it requires some feature engineering (word analysis, naively frequencies and manually reducing the text input to set of relevant words - or more advanced natural language processing).

This problem is for students who have more ML experience from outside this class and are looking for a larger challenge. You get no bonus points for attempting this other than bragging rights and experience working with slightly more complex data.

The data is from the 'Buy my a Pizza' Subreddit where Reddit users can make a case for other users to buy them a pizza. The data is several features about the request, along with the text of the request itself. Each request is labeled with whether or not the requesting user was given a free pizza.

The goal here is two-fold:

1. Predict which requests will actually receive a pizza.

2. What features about the requests are most predictive (i.e., if I wanted to make a request for free pizza, what are the things I should do to maximize my chances for a free pizza?).