Final Project

Points: 100
Team-based? OPTIONAL (no more than 2 students if you choose to do so)
Due: Wed, 2019-May-06

A portion of your final grade in CSCI 349 consists of the final project. This project will be a relatively large effort (compared to your work to date) to apply what you have learned in this course to the mining of a real-world, large-scale data set. The project normally requires 2-3 students per team, but given the circumstances this semester, teams will be optional. The overall constraints and goals for each project will be identical, and graded using the same rubric.

The final project deliverable will yield **Python Notebook files** pushed to your Git repository representing different phases of the project, including one final notebook presenting your final data preprocessing and model selection and performance results. **The final notebook is the primary deliverable that will be graded. The intermediate notebooks are used to help ensure you are making progress.**

1 Project ideas

You are to select a challenging dataset of significance and relevance. (Common datasets in the machine learning community, such as iris, titanic, MNIST, etc. are NOT acceptable.) Data size must be within the resources that can be handled by your computer. Alternatively, cloud-based resources (e.g. Google Colab) are entirely acceptable, be sure to copy over the URL(s) to your cloud notebook(s) to your readme.md file in your own Git repo.

For 2020SP – This is an unprecedented time in your life. COVID-19 has redefined what "normal life" is for everyone. Rest assured, we will go back to normal at some point in our future. This is certainly not the first pandemic to ever occur, and unfortunately, it will not be the last. However, let's hope it's the last in your lifetime! It will certainly be a memorable experience for life.

Data - What "dataset of significance" will you choose? You could consider using this time to explore some data surrounding this pandemic. Though, clearly with an ongoing crisis such as this, the data are changing daily. Unstable, dynamic data can be quite challenging, yet very interesting to explore. Besides, I can understand if you prefer to do something that keeps your mind off of this! Therefore, I'll leave my normal list of ideas below, and simply encourage you to select some dataset that is challenging, relevant, and of interest to YOU! I'm generally open to whatever you want to explore here, as long as the dataset is challenging in some way. Again, you can NOT use any standard ML dataset used repeatedly for academic purposes (e.g. Iris, Wine, Titanic, MNIST, etc.).

Some ideas:

- Kaggle competition. Choose any significant dataset and compete for money, knowledge, or kudos.
- KDD cup – Go to https://www.kdd.org/kdd-cup . These are significant datasets that have been used for ACM KDD competitions.
- UCI data repository (http://archive.ics.uci.edu/ml/) has one of the oldest, and still updated data repositories used for helping researchers who are studying machine learning and data mining. Some of the data here are quite challenging, and would make for an interesting project.

- **Google Dataset Search** - https://toolbox.google.com/datasetsearch - Are you looking for an interesting dataset related to a relevant problem of interest? **This is a GREAT place to start!**
- Investigation of a real-world problem / challenge – Suppose you have some area of interest to you. It could be healthcare, cancer, bullying, hate speech, climate change, astronomy, disease, colds, flu, pandemics, vaccine use, etc. Here, you and your team will focus on a problem of significance, and use data to help us understand the problem. It can be a social problem, an economic problem, a scientific problem, medical or healthcare problem, or any problem of significance today. There are no lack of data! You have hundreds, if not thousands of repositories worldwide, depending on the problem you are looking into.

  Some examples

  - Social media – seek datasets that are being used by the research community to help us identify posts that are "fake news", or hate posts, or posts from bots. Or, develop a model that can predict the sentiment of a post (happy? Sad? Hateful? Angry? Depressing? Etc.)
  - Climate change – there are no lack of data here. However, your challenge will be that you will need to learn how to deal with modeling and predicting from time series data.
  - Any human health related problem, including predicting some class from medical images, genetic / genomic data, EHRs, etc.
  - Astronomy – there are some fantastic (and HUGE) datasets that are being made public that are using imagery to identify exoplanets.

You should expect to spend time on every important step in data mining, including **preprocessing**, model evaluation and knowledge extraction. You might choose one method, and explore methods for improving the performance results through careful exploration of model parameters, and/or choose several classifiers and compare and contrast the performance of each of them. Expect to preprocess your data to get it in a form suitable for mining, and for many problems, expect this to take some time. Keep track of everything you do. **Use your Python Notebook files to tell your story and track your progress through your project!**

You should expect to use standard performance metrics to evaluate your models. You should not be blindly and randomly trying methods out! Justify your choices. Compare and contrast your results among the classification methods you choose, and clearly justify why each is performing the way it is. Many methods and/or datasets have supporting publications. Be sure to reference these papers, and report their results.

2 Timeline

For this semester, I am not requiring any hard timeline, only the final deadline is absolute. I suggest that you follow your path to completion by not waiting until the last minute. Here is my suggested timeline:

Step	% of grade	Date	What is due
1			Gitlab setup
2			Data selection
3	25%		`DataPrep_EDA.ipynb`
4	25%		`Modeling.ipynb`
5	50%	May 6	**`Final_Report.ipynb`**

3 Project steps

This section will lay out the details for each step required to complete the project, assuming you have selected your dataset of interest.

3.1 Gitlab Setup

For 2020SP: If you are completing your project alone, then create a new directory in your course repo called **`project`**, and keep your work there. Keep your data files in a subdirectory in project called **`data`**. If you are working with another member, then you can choose one member who will manage the deliverables, or follow these instructions below to have a new shared git repository specifically for the final project. PLEASE UPDATE YOUR **`Readme.md`** file as described below for team projects.

For those completing their final project as a team - create a repository on Gitlab (gitlab.bucknell.edu). Name your repo **`csci349_FinalProject`**. Add me (the instructor) as a **Developer**. I will receive an automatic e-mail from Gitlab notifying me that I've been added as a Developer to your project. **All work for your project will be delivered in Gitlab**.

NOTE: If you are comfortable working with branches in Git, that is fine (and desirable for team projects, not necessary for individual projects.) However, when you push your work to the remote repository, I will only pull from the **`master`** branch!

If branching, BE SURE TO MERGE YOUR LOCAL WORK TO THE `master` BRANCH, AND COMMIT WITH AN APPROPRIATE MESSAGE INDICATING THE MILESTONE, AND PUSH THE `master` BRANCH TO THE REMOTE REPOSITORY!

Be sure to create the **`Readme.md`** file summarizing the project, and please keep it updated throughout your project. This should be a brief summary of your project by the end.

Manage your data in a subdirectory called `data`. If your dataset is small (i.e. < 1MB in size), then you can commit your entire dataset into Git. However, if it is extremely large, you don't' want to commit this. You should commit only a small subset of your data so that I can test out your work when grading time comes.

Include a URL to all data you are using on the `Readme.md` page.

3.2 `DataPrep_EDA.ipynb`

Now, create your first notebook file, `DataPrep_EDA.ipynb`. Use both markdown and code cells to convey the following:

- What problem are you working on? Summarize in a single cell.
- What data are you using to understand the problem? Describe the data in a very general sense. Where did it come from? You should understand what every observation in the data represents, and what each variable represents.
- Remember that the key to achieving good data mining outcomes is understanding how each real-world entity in your data will be represented as a fixed length vector of attributes in your dataset! And, as we've repeatedly discussed in class, *preprocessing* your data will be a big part of this challenge. (This cannot be underestimated. If you do not expect to spend quality time cleaning and prepping your data, you will not get good results.) Once you have established how

each data object is represented in a form ready for a data mining algorithm, and the data are clean, you will have a substantial part of your battle toward modeling solved.

- You should strive to generate good summary statistics, show what the data looks like, and include good EDA and visualizations with boxplots, barcharts, density plots for key variables, or whatever other plots you want that are specific to your data and problem to help the reader understand basic distributions of important variables. Remember - visualizations can help you convey general info about your data and are extremely helpful.
- In your final cells, discuss the modeling methods you *expect* to use. Start by clearly explaining if this is a classification, regression, clustering, or association rule mining problem? Justify.
    - We have not covered *every* possible method out there. But, you have much of the framework to apply most algorithms, even those beyond what we covered in class. Feel free to explore different methods if you have good justification for doing so.
    - If there are any papers of significance that have been published with these data, then discuss the ones most interesting / relevant to the team.
- Finally, what is your overarching aim with this project? What are you hoping to learn? Or, what hypothesis are using the data to confirm or disprove? What challenges do you foresee on this project? Discuss your concerns. How will you get your work done? Give a reasonable list of milestones to reach to arrive at the final deadline for the project.

3.3 Modeling.ipynb

Copy over the important cells from the previous step that read in and cleaned your data to this new notebook file. You do not need to copy over all of your EDA and plots describing your data, only the code that prepares your data for modeling. This notebook is about exploring the development of predictive models. Some initial preliminary work on applying some modeling techniques should be completed

Be sure to commit and push all supporting code that you've completed in this NB file. Include in this notebook a summary cell at the top that details your accomplishments, challenges, and what you expect to accomplish for your final steps. Be sure to update your `readme.md` in your repository.

3.4 Final_Report.ipynb

This is your final report! Consider the scenario of you giving a report to your supervisor on these data. Take the parts from the previous two notebooks that help make this a complete story from start to completion. Include good reporting techniques. THIS NOTEBOOK SHOULD BE A COMPLETE STANDALONE NOTEBOOK FROM START TO FINISH! But, ONLY INCLUDE THE BEST OUTCOMES FROM YOUR PREVIOUS NOTEBOOKS! Your final report should only include your best visualizations from your EDA, your best model, and best performance results.

Include the following sections:

1. Introduction

   This should have mostly been done in your first notebook. Just copy over relevant cells from your first part of the project, and add any new information you have learned. Your aim is to motivate the reader with the importance and relevance of your project.

2. Data

Again, most of this was likely done in your first part of the project. So, feel free to copy the important cells over.

Introduce the original, raw data. Where was it collected? When? How? Explain the meaning of the variables. What does each observation represent? And be sure to explain the key target variable (assuming you are doing classification / regression)

Also, is there any related work? Identify any related work that have used these data. What methods did they use? How did they measure performance?

3. Data Preparation

   Again, most was probably done in the previous two sections. So, copy over what you are happy with, and be sure it's clean with your latest code that cleaned and prepared your data for modeling. Size, variable types, missing data, etc.

   Preprocessing steps should be explained, and explain why you did what you did. What did you need to do to clean and prepare your data for modeling? Include any dimensionality reduction techniques.

   Visualizations after preprocessing will do far more to convey your summary statistics, EDA, distributions, correlations with the target variable, etc. than just showing me lots of numeric tables

4. Model

   What methods did you ultimately determine was the best? What parameters were selected? Justify the selection of parameters. (i.e. did you do a grid search? how many different models did you choose? You can not simply say, "XGBoost was the only one I evaluated, and I used default parameters." You are expected to evaluate different models and different hyperparamters.)

5. Performance Results

   Clearly convey the results of your model. I expect to see ROC curves, precision / recall curves, confusion matrices mapped using matplotlib, tables with prediction performance by class, (or, if regression, use appropriate regression measures.)

6. Discussion

   Don't spend a lot of time here. I simply want you to reflect on your project. For example: discuss any challenges you had with cleaning and preparing the data. Did you find any surprises during your modeling? Compare and contrast the methods and hyperparameters you evaluated. And, it's often useful to discuss the features that you thought were the most predictive, and those that were least useful. (Search for feature importance scikit-learn for more info!) Any info that might be of interest to me related to your project goes here.

7. Conclusions

   A short summary of your findings goes here.

Be sure to make any final edits to `readme.md` that summarizes your findings.

Kaggle competition?

If you are competing in a Kaggle competition, you must still document everything you are doing in your own notebook file. Include results of your final standing in Kaggle if you were performing in a competition.