BA476 - Machine Learning for Business Analytics
Team project description

Instructor: Gerdus Benade (benade@bu.edu)

This project involves acquiring a data set you find interesting, describing it, and formulating and solving a prediction problem. In other words, you will complete a supervised learning project where the goal is to create a model that makes as accurate predictions in your domain as possible. You will have roughly 10 weeks to complete the project. The steps involved in this process are:

1. Obtain data you find interesting (see below for potential sources)
2. Clean and format the data
3. Iterate between the next steps until you are satisfied
   (a) Train and tune several types of models
   (b) Investigate the predictions of your models and add features as needed
4. Select and evaluate your final model

You should expect to spend ∼40% of your time on cleaning the data, ∼40% on iteratively improving your model and roughly 15% on parameter tuning.
Groups will consist of 3-6 students and will be published no later than 2 Feb.

1 Milestones and deliverables

There are two key milestones for the project:

1. 28 Feb: Teams present their project proposals in class.
2. 29 Apr, 1 May: Teams do their final presentations in class and submit the slide deck, data, and code.

It is natural for your problem to change as you work on it. However, if it changes substantially at any point, please consult with me. Distilling complex topics to their essentials is a skill. Do not exceed the time allocated to your presentations. You should be able to answer questions about your data and methods during the presentation.

1.1 Phase 1: Exploration and Proposal

During the first phase of the project, leading up to the proposal, your goal is to find a data set online and formulate an interesting prediction problem around the data set.
The proposal presentation (5 mins, 2-3 slides) should consist at least of the following

1. Clearly state the problem you are solving?
2. Explain what data sources you are going to use.
3. Specify your outcome variable and the predictors used to predict this outcome.

4. Describe your dataset: how many rows/columns? What types of variables? You can use the descriptive analysis lab as guide and tweak it to suit your needs.
5. Visualize a couple of interesting relationships between some of your predictors and the outcome.

The project proposal is developmental, you will not be graded on it. (Of course, failing to do a project proposal will be penalized.)
You should avoid:

- Time series prediction/forecasting: Avoid date where the main predictor is the outcome variable at some earlier time (eg. stock prices, where today's price is a great rpedictor for tomorrow's price).
- Small datasets: too few rows (say, less than 200) or columns (say, less than 10).
- Data sets with a lot of missing values.

1.2 Phase 2: Predictive modeling

I suggest you start working on the project as soon as it is approved. Generally, I expect all teams to (at least) try all the types of models that we discussed in class. If you don't try one of the models we discuss you should explain why. This phase concludes with a final presentation.

1.2.1 Mid-semester checkpoint

We will have a project workshop in class on 8 Apr. By this time in the semester, I strongly suggest that you have completed the following components of the project:

1. Imported and cleaned the data
2. Performed basic descriptive analytics
3. Trained a couple of models, including one regularized and one tree-based model.

Meeting these goals will help ensure that you can make proper use of the in-class time and have enough time to focus on the iterative process of tuning models, investigating its predictions (and errors), and using your domain knowledge to add predictors evaluating if your changes improved things.

1.3 Final presentation

For the final presentation (10 mins) the points below can guide your presentation. Emphasize those aspects that were not already discussed in the proposal.

1. State the problem.
2. Tell us who cares about this problem and why.
3. Describe you data – where it came from, what it contains, did you have to clean it, did you add features etc.

4. Present some interesting descriptive analyses (plots/tables).
5. Discuss the predictive methods you used. Why did you choose these, how did you tune the parameters? Which model was most accurate? Why?
6. Tell us about obstacles. What was the biggest challenge you faced and how you overcame it (or not, not every problem has a solution).
7. Conclude. What did you learn that can be put to practice?

You will be required to submit three things: (1) the slide deck (2) any python notebooks used for the project; (3) your data. Your notebooks should, at minimum, contain the code used to obtain the results and figures that you presented, though I expect that it will include additional analyses.

2 Data

A link to a spreadsheet will be posted on the course page. Once you have settled on a problem, please post there so that we can avoid duplicate projects.
You can use any (non-proprietary) data. The following sources may be useful:

- Open Gov data: `www.data.gov`, `www.data.gov.uk`
- UCI ML Repository: `http://archive.ics.uci.edu/ml/index.php`
- Kaggle: `https://www.kaggle.com/datasets`
- KDD Nuggets: `http://www.kdnuggets.com/datasets/`
- NY City: `https://nycopendata.socrata.com/`
- Yahoo: `https://webscope.sandbox.yahoo.com/`
- Census: `http://www.census.gov/data/developers/data-sets.html`

3 Grading

The following is a rough guideline of how the project will be graded.

10% Relevance and Usefulness. Should this question be solved using machine learning methods? Did you solve it using methods we learnt in class? Is this a question worth answering/does anyone care?

25% Appropriateness and understanding of the data. Can you asnwer the question using this dataset? Do you understand why the data exists? Are there enough instances? Has it been properly cleaned? Did you demonstrate an understanding of the space through feature engineering? Did you incorporate outside data sources where needed?

35% Soundness and quality of the experiments. Did you try the obvious approaches? Did you try multiple methods and attempt to tune parameters? Did it work? If not, do you understand why?

10% Interpretation of results. What is the impact of your project? Do you understand its limitations?

20% Clarity and flow of presentation.

3.1 Grading baseline

You will find below a rough baseline used when grading. Exceeding (or failing to meet) a benchmark in a category will lead to an upwards (downwards) adjustment for that category.

3.2 Peer evaluation

Towards the end of the semester students will evaluate their fellow group members. The individual grades of group members will be adjusted based on contributions to the group (both positive and negative), up to two letter grades.
In well-functioning groups where everyone participates positively the most likely outcome is that everyone receives the same grade.

Criteria	Max	Score	Comments
Relevance/ Usefulness	10	9	• The problem is obviously a prediction problem. Correctly treated as classification/regression. • Not much of a stretch to see your model used in practice.
Data handling	25	20	• Data was cleaned appropriately. • Basic imputation of missing values. • Little to no data from external sources used to augment original data. • Minor feature engineering - new features are unconvincing or don't affect predictive accuracy.
Experiments	35	30	• Tried all (relevant) models discussed in class. • Little to no investigation of predictive models not explicitly treated in class. • All parameters are tuned using CV (but not nested CV). • Little to no use of GridSearchCV for models with multiple parameters.
Interpretation	10	8	• Brief discussion of implications/applications of model.
Presentation	20	16	• Detailed descriptive/exploratory analysis to highlight trends and quirks in data. • Visualisations are easy to understand • Clear baseline performance established • Shows effect of parameter tuning for selected models. • Final train/test performance shown for all models. • Little to no quantification of impact of each component (data processing/imputation/feature engineering etc.) • Limited discussion of model's generalization.
Total	100	83	