Assignment 3

CSCI 3141: Foundations of Data Science Using R

Mariana Oliveira

November 10, 2022

Purpose

Predictive Analytics comprise a set of statistical and machine learning tools that allow you to make accurate predictions about future or unknown instances or events based on historical data.

This assignment will give you the opportunity to practice building a data science solution to a predictive (classification or regression) problem. You will train and test some of the learning algorithms you have learned about in class, make important decisions about how to evaluate your own solutions, and get a deeper understanding of the challenges posed by prediction using real-world data sets.

The final estimated performance you get from your methodology is not nearly as important as the **process** you will go through to get there. The main objective of this assignment is not actually to get the best possible performance out of your models. Your main goal is a much more pedagogical one.

The aims of this assignment will be to identify and define a predictive task, choose appropriate performance metrics and estimation methods, carry out experiments using different models and appropriate pre-processing methods, and correctly interpret and reason about the results. The task instructions as well as the grading rubric will guide you in how to achieve this.

Task

In this assignment you will use R markdown to create a report of a predictive data science solution for your data set of choice.

You will be able to build upon your work from the previous assignment. Having explored your data before, the insights you have gained may help you make decisions about how best to use your data to improve the predictive capacity of Machine Learning models.

In this assignment, you should go through the following phases:

1. Data Understanding:
   - You should get acquainted with your data set, understand the its shape and format, the variables you have available, identify data quality issues, etc. This should be a quick step if you already completed Assignment #2 with the same data set;
2. Problem Definition:
   - Since you do not have a client requesting you to solve a specific problem, you might need to read more about how your data has previously been used in a predictive analytics context;
   - Identify and formulate the predictive task you will be tackling, and be sure to familiarize yourself with your target variable. If you are not sure about this step, contact me (or the TA, if the strike is no longer ongoing) as soon as possible;
3. Experimental Design (and Project Planning):
   - Select an appropriate performance metric and performance estimation method given the characteristics of your predictive task and target variable distribution (These two decisions are probably the most important in this assignment, aside from problem definition, and they should be made before

   starting to work on the rest of the project. Note that you will need to justify these decisions in your report);
   - Think about the experiments you would like to carry out. You should consider the following:
     - What can you use as a baseline result to compare your solutions against?
     - Which learning models do you want to try? Will you try them in a specific order and/or if you find specific properties in your data?
     - Will you be experimenting with different hyper-parameters with one (or more) of the models? How will you go about that?
   - Remember the timeline for this assignment and try to make a plan for how you will execute it;
4. Data Preparation:
   - Review your data quality assessment from the previous assignment;
   - Consider the learning algorithms you will be using and prepare the data set so it will be ready for modelling. You should consider what you have learned in class so far:
     - Will each model work with the type of features you have available?
     - Do you need to take steps to handle missing data and/or outliers?
     - Does the model require some other kind of pre-processing or data transformation to work well with your data?
     - Should you be adding additional variables (or discarding others) given your knowledge about your data set, or domain knowledge you have about the topic?
   - Remember to document all the pre-processing steps you take, and justify your decisions;
5. Data Modelling and Analysis:
   - This is where you run your experiments, compare and analyse your results (including conducting statistical significance tests);
   - Based on your analysis, you should be able to make recommendations about how to address your predictive task (e.g., pre-processing steps needed, best model and model hyper-parameters to use);
   - Remember that you can always go back to the previous stage and prepare your data better to try to improve your results (Data Science is rarely a linear process);
6. Reporting: Polish your final derivable and make sure it follows the guidelines.

Prioritize your work: it's important to **have a working "skeleton" solution early on** (e.g., using a single learning algorithm) that you can improve upon afterwards. If you have a pipeline implementing the several steps in your experiments using a single learning model, you can then adapt it to work with different models or model variants. You can always run additional experiments later if you have the time (e.g., trying out different hyper-parameters for a specific model).

Keep in mind the main pedagogical goals of the assignment discussed above: while it is nice if you can get very good predictive performance, it is more important for you to carry out technically sound experiments, properly analyse your results and justify the pre-processing steps you took, and modelling decisions you made.

Document your process as you go along. That will make it a lot easier to create your final report.

Deliverable

The main deliverable for this assignment will be a report in the form of an R notebook. You will have to submit both the .nb.html and the .Rmd files as well as any other files you used in a zipped folder.

You should use markdown headers to include at least the following four sections:

1. Introduction
   - Briefly present your data set, and describe the predictive problem you will be tackling;
2. Experimental Setup
   - Describe your experimental setup, including the evaluation metric and performance estimation method you selected. Justify your decisions;
3. Results
   - Present the results of your experiments. This should include:
     - Tables and/or visualizations showing the results obtained with different methodologies (i.e., with different learning algorithms, algorithm hyper-parameters, and/or data preparation

methods);
    - Model comparison using statistical significance tests;
    - Brief textual descriptions interpreting the results;
4. Conclusion
    - Share your main conclusions and provide recommendations for addressing your predictive task;
    - This should include a recommendation about which learning model (and model hyper-parameters) you would select for this task and your reasons for it (including reference to statistical significance tests).

Guidelines

- Cite any source you used (article, book or web page) or consulted and make explicit which part of the work it influenced;
- Clear the output and re-run all your code chunks early and often (ideally, every time you include a new R code chunk). This will help you detect issues early on. The .Rmd file alone will not be acceptable as a deliverable;
- Use RStudio to spell check your report;
- Re-read your report, edit for clarification and remove duplicated information.
- Remove superfluous code and output (e.g., printing a data set to the screen);
- Look at the grading rubric at the end of this document to help you decide the level of detail required.

This assignment should be completed **individually**. You can discuss your assignment with your colleagues, but do not share your code or report. If you have any doubts about how to proceed, you can review the Academic Integrity module in the course at: https://dal.brightspace.com/d2l/le/content/230452/Home.

Data

Ideally, you should use the same data set you used in your previous assignment, so that you can take advantage of the knowledge you gained by performing EDA on your data set. If there is some reason you would like to change your data set or if you are having trouble defining a prediction problem using your data set, please **contact me** (or the TA, if the strike is no longer ongoing) about it as soon as possible.

Criteria

The assignment will be graded according to the following criteria:

- Appropriate framing of prediction problem and description of experimental setup;
- Appropriate choice of performance metric and performance estimation method;
- Adequate data preparation;
- Diversity of modelling algorithms, and tuned hyper-parameters;
- Adequate model comparison and analysis of overfitting and interpretability;
- Appropriate presentation of results and conclusions;
- Well-written and organized report, clear and concise;
- Bonus marks for exceptional work showing technical creativity, and for draft submissions of the assignment (see details on Brightspace).

A detailed rubric is presented below. **Reading the rubric** can help you better understand what will be expected of you in this assignment.

Rubric

Criteria descriptions from novice (lowest mark) to proficient (highest mark).

Introduction and Experimental Setup Description

- Data and problem description essentially copied from the data source and/or the previous assignment. Unclear which set of experiments will be carried out
- Description of the data set that is superficial or incomplete, or goes into too much detail. Superficial description of the prediction problem. Description of experimental setup may be somewhat confusing.
- Concise description of the data that makes it clear what data is being used and where it was obtained. Clear definition of the prediction problem that will be addressed. Clearly explained experimental design, aligned with the data science goals of the assignment.

Evaluation Metrics

- Inappropriate choice of performance metric given the problem at hand, with no attempt of justification.
- Chose and correctly calculated a basic performance measure (e.g. accuracy, error, MSE), with little justification.
- Chose and calculated appropriate performance measure(s) with proper justification given the characteristics of the problem. May have used more than one metric (e.g., precision and recall) to better understand model performance.

Performance Estimation Method

- No train-test split.
- Training and testing on separate data, but with leakage.
- Training and testing on correctly separated data.
- Training and testing on correctly separated data, with multiple splits and/or nested evaluation for model selection.

Data Preparation

- Raw data set used directly, with little to no preparation. No new features added or irrelevant features removed.
- Simple data preparation transformations such as central imputation of missing values. Added simple new features (e.g., ratios) and/or removed unhelpful features (e.g., features with too many missing values). Decisions sometimes not sufficiently justified or motivated by understanding of the modelling algorithms used.
- Appropriate data preparation steps taken, justified by understanding of the selected modelling algorithms. Added or removed features based on model performance estimates and/or statistical significance tests. May have experimented with more complex statistical feature engineering such as PCA. May have used more advanced data pre-processing such as feature standardization or one-hot encoding employed in useful ways.

Predictive Modelling: Diversity of Algorithms

- Single learning algorithm tested.
- Tested less than three algorithms with significantly different functional form assumptions.
- Tested three or more learning algorithms with significantly different functional form assumptions.

Predictive Modelling: Hyper-parameter Tuning

- No hyper-parameter tuning. Only default values used.
- Limited, manual parameter tuning for at least one learning algorithm..
- More systematic parameter tuning (e.g., using grid search) of at least one learning algorithm. Alternatively, limited, manual parameter tuning of at least two learning algorithms. May have unclear understanding of the effect of each parameter on the learning algorithm(s).
- Parameter tuning of at least two learning algorithms, where at least one was systematic (e.g., using grid search). Shows clear understanding of the effect of each parameter on the learning algorithms.

Results: Model Comparison

- No baseline to compare model performance against. Statistical significance tests missing.
- Baseline included, but may be unsuitable. Conducted statistical significance tests, but their interpretation may be lacking.
- Appropriate baseline included. Statistical significance tests conducted and interpreted correctly.

Results: Analysis of Overfitting and Intepretability

- Missing or incorrect analysis of overfitting and feature importance or model interpretability.
- Superficial analysis of overfitting, with no attempt at addressing it. Some analysis of feature importance and/or model interpretability.
- Correct analysis of overfitting, and made an attempt at addressing it. Analysis of feature importance and inspection of "white-box" models.

Results: Presentation (graphs, tables, textual descriptions)

- Missing tables and/or figures summarizing the results.
- Results presented using tables and/or figures that convey information but lack context for interpretation.
- Results presented clearly and concisely, with the help of visualizations and tables whenever appropriate, as well as brief textual descriptions.

Conclusions

- Recommendations missing, incorrect, or not based on analysis.
- Recommendations only partially correct or complete (e.g., missing analysis of statistical significance).
- Well-justified model selection given the analysis of results and context of the problem.

Report (and Code) Readability

- Only R code or output present, and no discussion of results. One or more major issues that make the report difficult to read or understand (e.g., long output dumps, inappropriate or no use of Markdown headers, presence of extensive code irrelevant to the analysis).
- Results and discussion present but not following the suggested report format. A few minor issues with the R code (e.g., some irrelevant code shown, a few variable names that could be more readable, one or two places that could benefit from comments, or overly commented code). Some large grammar and/or spelling issues.
- Well-written, clean and well-organized report. Clear and concise. Using full sentences and spell-checked. Follows the suggested report format and uses Markdown headers appropriately.

Bonus

- Early Submissions:
  - For the first draft, you should at least have completed stages 1 to 3 of the assignment and started working on stage 4.
  - For your second draft, you should already be on stage 5 of the assignment and have a "skeleton" solution (though you can still run some additional experiments afterwards).
- Creativity & Originality: Exceeded the parameters of the assignment, showing technical creativity beyond what was taught in class.