Table of Contents

[Homework 2 1](#_Toc118382210)

[Finding Data 1](#_Toc118382211)

[Report Submission Instructions: 1](#_Toc118382212)

[1. Problem Definition (5 marks) 1](#_Toc118382213)

[2. Exploratory Data Analysis (25 marks) 2](#_Toc118382214)

[3. Data Treatment (10 marks) 2](#_Toc118382215)

[4. Model Development and Tuning (10 marks) 2](#_Toc118382216)

[5. Model Evaluation (10 marks) 2](#_Toc118382217)

[6. Code Appendix (10 marks) 3](#_Toc118382218)

[Restrictions 3](#_Toc118382219)

# Homework 2

Perform machine learning predictive algorithm for with any data set that interests you.

## Finding Data

You can find data at:

* <https://data.world/search>
* <https://www.kaggle.com/datasets>
* Web scraping
* API’s

## Report Submission Instructions:

Due the final class lab day before exams.

Please submit either a word document or a Jupyter Notebook. If you submit a Jupyter Notebook please ensure the data remains in your Notebook. Do not use pdf please. Also, please include your dataset with the submission.

Please submit a report which contains the following:

### Problem Definition (5 marks)

Explain the problem you are trying to solve.

### Exploratory Data Analysis (25 marks)

Please provide a thorough exploratory data analysis. Make sure the report is informative and rich with information for a non-technical audience. Use visualizations. Be creative, professional and please show more than just a heat map or a list of variables with data types.

#### (EDA) Feature Segmentation

Categorize the X feature set into two to four groups if you can. You may use factor analysis to help. Explain how each of these categories are different. For example, segment A is likely to say ‘yes’. Segment ‘A’ has attributes that look like …**.** Segment B is likely to be indifferent. Segment ‘B’ has attributes that look like …**.** Segment C is likely to say ‘no’. Segment ‘C’ has attributes that look like …**.**

Note: If you use PyCaret please do not just copy and paste the output from PyCaret into your report. I would like to see careful thought put into the narrative and a conscious effort to remove unnecessary information.

### Data Treatment (10 marks)

In this section discuss how you prepared the data for modelling. List any variables that you created or treated and explain the modifications.

### Model Development and Tuning (10 marks)

Use the following automated feature selection routines and compare the results in a reader-friendly chart which shows variable selected by each:

* Recursive feature elimination
* Forward feature selection
* Feature Importance (See notes for random forest)

Use this section to explain how you created and tuned your model. You are required to build the following types of models:

* A neural network.
* A bagged model.
* A stacked model.
* A stand-along machine learning model.
* Linear regression or logistic regression.

### Model Evaluation (10 marks)

Present statistics and summaries that compare your models in a professional user-friendly manner for a non-technical audience to explain how your model was evaluated and selected.

### Code Appendix (10 marks)

Please place your code in the appendix. Keep it neat and tidy. Use a mono-spaced font. Or, you can use a Jupyter Notebook.

### Restrictions

Do not use data sets that are covered in the class. Please do not use the same type of model that you use for any of your other assignments in COMP3948, COMP4948 and COMP4949.