Table of Contents

[Submission Instructions 1](#_Toc94076375)

[Requirements 1](#_Toc94076376)

[TrainingScript.py (20 marks) 1](#_Toc94076377)

[EnsembleScoreScript.py (15 marks) 2](#_Toc94076378)

[Report 2](#_Toc94076379)

[1. Problem Definition (1 mark) 2](#_Toc94076380)

[2. Exploratory Data Analysis (10 marks) 2](#_Toc94076381)

[3. Data Treatment (3 marks) 2](#_Toc94076382)

[4. Model Development, Tuning and Comparision(6 marks) 2](#_Toc94076383)

[Competition 2](#_Toc94076384)

[Deductions 2](#_Toc94076385)

Using the **ds\_salaries.csv** file you are to build a stacked neural network to predict data scientist salaries **salary\_in\_usd**.

# Submission Instructions

Please submit a single **.zip** archive (no .rar please) contains these items:

* **DS\_Salary\_Report.docx (no pdf’s please)**
* **TrainingScript.py**
* **EnsembleScoreScript.py**
* **1 BinaryFolder** (which contains all binary ANN models and binary scalar(s))

# Requirements

Please meet the following requirements:

### TrainingScript.py (20 marks)

* Runs continuously from start to finish without error.
* Loads and prepares the data needed to build the models.
* Builds all models with significant variables.
* All stand-alone models must be artificial neural networks.
* At least optimizes the number of nodes, learning rate, number of layers, activation function and kernel initializer for the network model. **(13 marks)**
* Has more than one hidden layer.
* Saves your model binary files with early stopping. **(5 marks)**
* Saves your binary scalar(s) **(2 marks)**
* Uses a random 80/20 - train/test split.

### EnsembleScoreScript.py (15 marks)

* No training code in it or marks will be deducted. The stacked model cannot be fitted in this script. All fitting must take place in TrainingScript.py.
* Contains no data exploration code or marks will be deducted.
* Contains no model evaluation code or marks will be deducted.
* Prepares (and imputes where needed) data from a file that uses a format similar to **ds\_salaries\_mystery.csv**. This file may contain any number of rows.
* Loads all binary scalars to scale the data.
* Loads all binary models needed to prepare the data for the stacked prediction.
* Outputs proper number of predictions to a file in the same format as **ds\_salaries\_predictions.csv**.
* Runs continuously from start to finish without error.

## Report

### Problem Definition (1 mark)

Explains the problem being solve.

### Exploratory Data Analysis (10 marks)

Provides a reader-friendly, visually efficient and professional looking overview of insight about relevant data attributes and how they are related to the target.

### Data Treatment (3 marks)

Discusses how data was prepared for modelling. Lists any variables that you created or treated and explains the modifications.

### Model Development, Tuning and Comparision(6 marks)

Discusses how models were created and tuned. Discusses differences in how stand-alone models were built. Discusses how individual and stacked models performed and how they were evaluated. Include loss plots and RMSE in your discussion. Discuss how you accounted for variance of the results in a reader-friendly manner.

## Competition

To qualify for the competition:

* Your project must be submitted before the due date.
* Your project must be relatively easy to set up and run.
* Code scripts must run without error.
* Reasonable effort has been taken to prevent over-fitting.

## Deductions

* Does not save all scalars and trained models to the same binary folder.
* Does not load all scalars and trained models from the same binary folder.
* More than one variable is created to reference to the same binary folder.
* Complicated or confusing deliverables which are difficult to set up and test.
* Lack of professionalism.
* Above average RMSE’s or lack of signs of a reasonable effort to achieve decent predictive results.
* Code that does not run from start to finish without error.
* Rows with missing data are dropped rather than imputed. Dropping is not permitted.