Final Project – Self-Grading

Shradha Godse – Credit Risk

I Machine Learning Question: 20 pts

A. Is the background context for the question stated clearly (with references)? → 4

B. Is the hypothesis/problem stated clearly → 5

C. Is it clear why the problems are important? Is it clear why anyone would care? → 5

D. Is it clear why the data chosen should be able to answer the question being asked? → 4

E. How new, non-obvious, and significant are your problems? Do you go beyond checking the easy and obvious? → 5

II Data Cleaning/Checking/Data Exploration: 20pts

A. Did you perform a thorough EDA (points below included)? → 3

B. Did you check for outliers? → 3

C. Did you check the units of all data points to make sure they are in the right range? → 1 – multiple features, so went through a few

D. Did you identify the missing data code? → 4

E. Did you reformat the data properly with each instance/observation in a row and each variable in a column? → 5

F. Did you keep track of all parameters and units? → 4

G. Do you have a specific code for reformating the data that does not require information not documented (eg. magic numbers)? → 5

H. Did you plot univariate and multivariate summaries of the data including histograms, density plots, and boxplots? → 2

I. Did you consider correlations between variables (scatterplots)? → 3

J. Did you consider plot the data on the right scale? For example, on a log scale? → 0

K. Did you make sure that your target variables were not contaminating your input variables? → 4

L. If you had to make synthetic data was it a useful representation of the problem you were trying to solve? → 4

III. Transformation, Feature Selection, and Modeling: 30pts

A. Did you transform, normalize, filter the data appropriately to solve your problem? Did you divide by max-min, or the sum, root-square-sum, or did you z-score the data? Did you justify what you did? → 4

B. Did you justify normalization or lack of checking which works better as part of your hyper-parameters? → 5

C. Did you explore univariate and multivariate feature selection? (if not why not) → 5

D. Did you try dimension reduction and which methods did you try? (if not why not) → 5

E. Did you include 1-2 simple models, for example with classification LDA, Logistic Regression or KNN? → 4

F. Did you pick an appropriate set of models to solve the problem? Did you justify why these models and not others? → 4

G. Did you try at least 4 models including one Neural Network Model using Tensor-Flow or Pytorch? → 4

H. Did you exercise the data science models/problems we described in the lectures showing what was presented? → 5

I. Are you using appropriate hyper-parameters? For example, if you are using a KNN regression are you investigating the choice of K and whether you use uniform or distance weighting? If you are using K-means do you explain why K? If you are using PCA do you explore how many dimensions such as by looking at the eigenvalues? → 4

IV. Metrics, Validation and Evaluation 20pts

A. Are you using an appropriate choice of metrics? Are they well justified? If you are doing classification do you show a ROC curve? If you are doing regression are you justifying the metric least squares vs. mean absolute error? Do you show both? → 5

B. Do you validate your choices of hyperparameters? For example, if you use KNN or K-means do you use cross-validation to optimize your choice of parameters? → 5

C. Did you make sure your training and validation process never used the training data? → 4

D. Do you estimate the uncertainty in your estimates using cross-validation? → 4

E. Can you say how much you are overfitting? → 4

V. Visualization 10pts

A. Do you provide visualization summaries for all your data and features? →2

B. Do you use the correct visualization type, eg. bar graphs for categorical data, scatter plots for numerical data, etc? → 3

C. Are your axes properly labeled? → 4

D. Do you use color properly? → 5

E. Do you use opacity and dot size so that scatterplots with lots of data points are not just a mass of interpretable dots? → 4

F. Do you write captions explaining what a reader should conclude from each figure (not just saying what it is but what it tells you)? → 5

VI. Code 20pts

A. Is the code provided can reproduce the entire work? → 4

B. Is the data included or at least linked (externally) with instructions on how to download it? → 5

C. Do you factor repeated operations into functions to avoid repetitively and error-prone copy-paste? → 4

E. Do you use docstrings and numpy documentation style:

https://github.com/numpy/numpy/blob/master/doc/HOWTO_DOCUMENT.rst.txt

to make your code clear and readable? → 4

F. Do you use markdown cells to explain every step of your code similar to Homeworks and some example notebooks? → 5

G. Does the code demonstrate considerable work given the number of people on the project? → 5

VII. Presentation 30pts

A. Do you tell a coherent story with a beginning, middle, and end? → 4

B. Do you introduce why the problem is important? → 5

C. Do you explain in the first couple of slides what you accomplished on solving the problem? → 1

D. Are you careful not to have slides filled with text (keep in notes)? → 3

E. Is data and evaluations presented as clear figures (mostly)? → 3

F. Do you make sure to say what is or should be learned from each figure? → 3

G. Do you stay within your time limits 15 min? → 5

H. Do you avoid useless padding slides of no relevance? 5

VIII. Report 30pts → 4