# Title

Your name

### Air Force Institute of Technology

### *I have color-coded this template to show which sections are required for homeworks 4, 5 & 6. When turning in HW4, please change your updated HW4 text to black, and retain the placeholder text for HW 5/6. Similarly, when turning in HW5, retain the placeholder text for HW6.*

* Homework 4
* Homework 5
* Homework 6

### Abstract — Describe and motivate the problem. Mention your approach and list your key findings or results, along with why they are important. Include 7 or more numbers, such as number of datapoints, modeling metrics, or the benefits of the model. Compare the performance of the best model to the trivial model.

*Keywords — blah, blah and blah*

Style notes:

Please use the “IEEE para” style for each paragraph.

Be sure to use the “IEEE bullet” style below for any bullet points.

* a
* b
* c
* d

### I. Introduction & Background

Problem motivation

Mission understanding, and background needed to understand the problem. Include 2-3 references.

Your research question. This could be similar to “Using available data, how well can classical machine learning and neural network algorithms predict (describe output)?

Your success criteria, which could include performance or the ability to generalize.

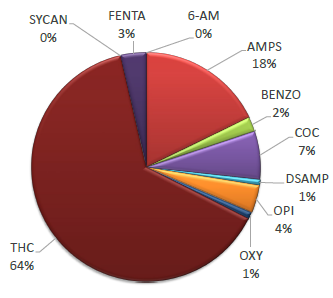


Fig. 1. (using “IEEE caption” style) AFDTL positive drug result distribution, March – June 2020 [2]. Unlike this example, be sure to define any acronyms in the caption, if they aren’t previously identified in the text.

Table 2. Prior machine learning analyses in XYZ topic area

|  |  |  |  |
| --- | --- | --- | --- |
| Description of work | Method used | Performance | Ref |
| Prediction of THC use in college students using personality trait data | Decision tree | AUC = 0.75 | [1] |
| Prediction of ecstasy use in military members using demographic data | Neural network | Precision = 0.60  Recall = 0.70 | [2] |
| Prediction of oxycodone abuse based on surgery type | Random forest | F1 = 0.60 | [3] |

1. *Data Acquisition*

A description of the process used to acquire your data, such as the originating office, survey or experiment.

1. *Data Understanding*

Perform relevant Data Understanding calculations and create visualizations, such as descriptive statistics, histograms or scatterplots.

* Include the most insightful 3 visualizations in your paper, along with an explanation.
* Include a table with the number of observations, features and distribution of your input and output variables.
* If you are performing classification, address the number of classes, and if the class representations in the data are balanced, or unequally represented.

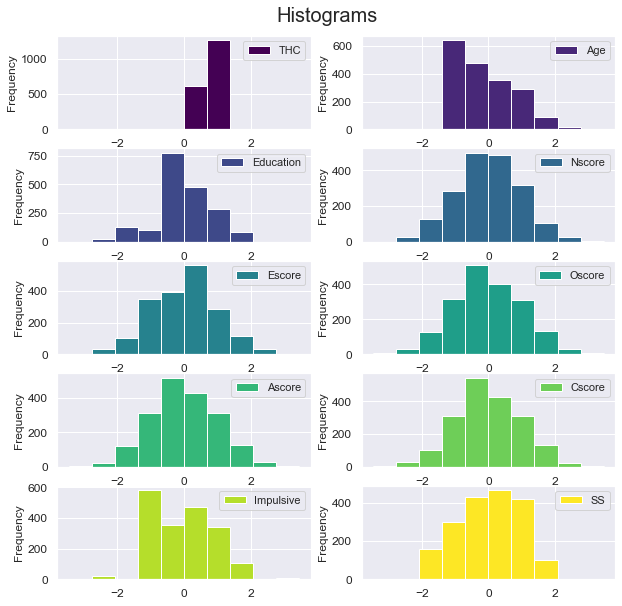


Fig. 2. (a figure idea) Variable Histograms.

Table 1. (a table idea) Feature Summary

|  |  |
| --- | --- |
| Input Variable | Data Distribution |
| Year of Launch  LV Type Flight Count | ~Uniform  ~Log-Normal |
| LV Family Flight Count | ~Log-Normal |
| Org Flight Count | ~Log-Normal |
| Country Flight Count | ~Log-Normal |
| Number of Liquid Stages | ~Normal |
| Number of Solid Stages | ~Log-Normal |
| Total number of stages | ~Normal |
| Number of boosters | ~Log-Normal |
| Length | ~Normal |
| Mass | ~Normal |
| Diameter | ~Log-Normal |
| Country of Origin | Categorical |
| Developer Experience  Vehicle Experience Level | Binary  Binary |

### II. Method

(use past tense for all, as if you have already done the work)

A few introductory sentences.

1. *Data Preparation*

Document the methods that you used, such as transformations or filtering. If applicable, include up to 2 before/after figures showing the results of your data preparation.

1. *Metrics*

This example is for classification: “In this analysis, the classification metrics of precision, recall and f1 are used to measure the performance of the classical and neural network models. In the dataset, the positive class is a food crisis. In the case of an actual food crisis, it is critical that a model prediction of no food crisis be avoided. As a result, recall is the most important metric, which helps to avoid false negatives. Precision was also selected, as accurate predictions of true positives are important, and f1 was chosen due to the unbalanced nature of the dataset where only 15.3% of labels are positive. These 3 metrics also facilitate a direct comparison to prior work. In certain cases, the area under the receiver operating characteristic curve (AUC) and accuracy are also presented to highlight differences between models.”

1. *Classical Modeling*

Justify your selection of modeling variations, and briefly describe the algorithms. Variations could include different algorithms, adding or removing input variables, or transformations.

Discuss the contributions of input variables to your model (z or p test).

Document how you monitor to prevent overfitting, such as a train/validate split, or cross-validation.

1. *Neural Network Modeling*

Justify your selection of optimization algorithm, loss function and regularization technique, based on the type of problem you are modeling (see concept map).

Document how you monitor to prevent overfitting, such as a train/validate split, or cross-validation. Justify your selection, based on your number of datapoints and number of input variables.

Discuss what hyperparameter sweeps you will perform, and what regularization technique you will explore.

### III. Analysis & Results

A few introductory sentences.

1. *Classical Modeling*

Document performance metrics for each modeling variation:

* R2, MSE and residual analysis for regression
* Accuracy, precision, recall and area under the ROC curve (AUC) for classification. For classification, if you have an unbalanced dataset it might be good to focus on the number of false negatives and false positives

Add a table that summarizes the metrics for your 3 variations.

* If regression, also calculate metrics that would result if your model always predicted the mean of the labels
* If classification, also calculate metrics that would result if your model always predicted “1” and if it predicted randomly (chance).

Justify your selection of the best model

Table 2. (example) Logistic Regression Model Results. (note to class: the P-value selection, RGE and Select K Best are just examples – you can choose the modeling variations that you try)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | P-val | | AUC | Precision | Recall | Accuracy |
| Full | | 0.00 | 0.58 | 0.73 | 0.18 | 0.86 |
| P-value Selection | | 0.00 | 0.58 | 0.81 | 0.16 | 0.86 |
| RFE | | 0.00 | 0.58 | 0.81 | 0.16 | 0.86 |
| Select K Best | | 0.00 | 0.59 | 0.75 | 0.20 | 0.87 |
| Chance | | -- | 0.50 | 0.15 | 0.50 | 0.50 |
| Always Predicts Crisis | | -- | -- | 0.15 | -- | 0.15 |
| Never Predicts Crisis | | -- | -- | -- | -- | 0.85 |
| Goal | | -- | -- | 0.85 | 0.75 | 0.90 |

1. *Neural Network Modeling*

Document model performance that results from a hyperparameter (HP) sweep on neurons per layer, number of layers and 2 other hyperparameters.

Investigate how 1 regularization technique affects your best model, and document the results.

Document performance metrics from your modeling similar to the classical section above. Include rows for:

* Preliminary NN model
* After HP sweep
* After regularization

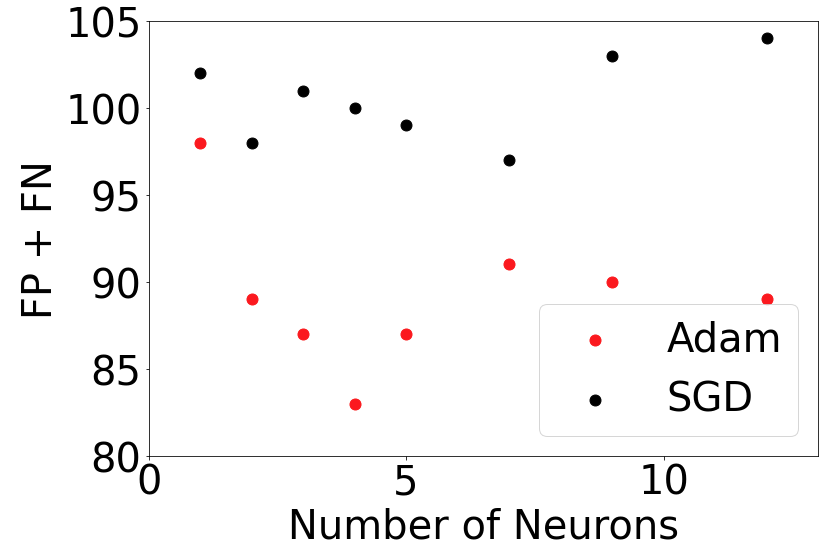


Figure 5. Model performance resulting from hyperparameter sweep.

1. *Model* *Evaluation*

Evaluate the results of your efforts using the criteria established in homework #4. Assess the value of this effort to your organization.

Discuss the final model selected, and why it is was selected.

Mention inferences that can be drawn from the modeling and from the data. Particularly insightful findings could be mentioned in your abstract and conclusion.

Include a figure that shows your model is not overfitting the data, or has an acceptable level of overfitting (<10%)

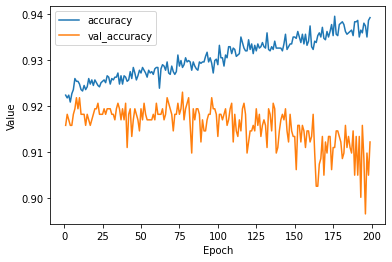


Figure 5. Neural network overfitting the dataset.

(student example) “With significant capacity built into the model, there is a significant divergence between the training accuracy and the validation accuracy. This model is clearly overfitting the data. A reduced capacity model was then implemented, and the number of epochs was reduced to 100. The results are shown in Figure 6.”

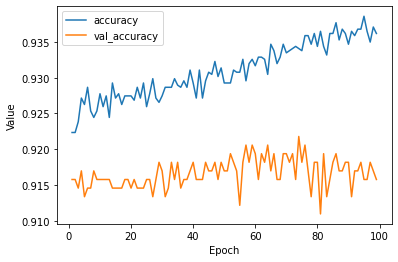


Figure 6. Overfitting corrected by XYZ.

1. *Model Application (a.k.a. deployment)*

Talk the audience through a simple application of your model.

How can it be used to further the mission of your organization? This could include:

* Qualitative applications:
  + mention how you could apply inferences from classical modeling - here is an example from an AF Drug Testing Lab student:
  + Inference “younger, less educated individuals who exhibit sensation seeking behavior and are open to experience tend to be at higher risk for THC use”
  + Application: we recommend a personality survey for new employees, and focus drug prevention efforts on those at highest risk for TCH use
* Quantitative applications:
  + how can you best use the prediction of the model to make a decision?
  + You could create an excel (or Python) analysis showing how a decision could be made
  + Be sure to mention the impact of that decision (i.e. we could save $500K, avoid 20% of space launch failures, or predeploy food aid to areas that have the highest risk for a food crisis

### IV. Conclusion

Restate main points, and avoid adding new information.

### References

Note to instructor: I am using the XYZ bibliography tool (please replace XYZ with Word, Mendeley, Zotero etc)

|  |  |
| --- | --- |
| [1] | G. Chen, Z. D. Dong, D. J. Hill, G. H. Zhang and K. Q. Hua, "Attack Structural vulnerability of power grids: A hybrid approach absed on complex networks," *Physica A: Statistical Mechanics and its Applications,* vol. 389, no. 3, pp. 595-603, 2010. |
| [2] | E. W. Prehoda, C. Schelly and J. M. Pearce, "US strategic solar photovoltaic-powered mircrogrid deployment for enhanced national security," *Renewable and Sustainable Energy Reviews,* vol. 78, pp. 167-175, 2018. |
| [3] | M. Ahmad, Operation and Control of Renewable Energy Systems, Hoboken: Wiley, 2017. |