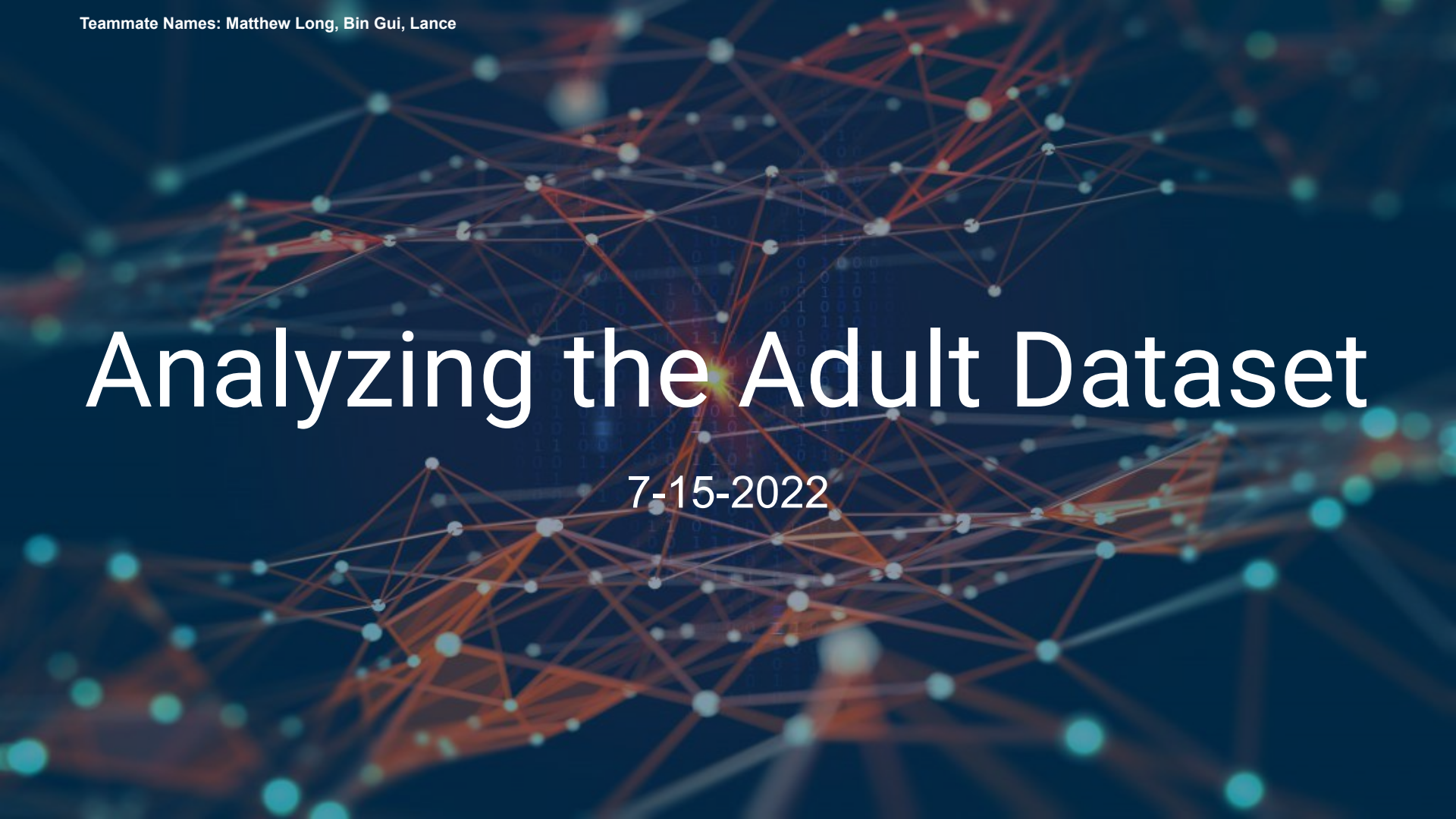
Teammate Names: Matthew Long, Bin Gui, Lance



# Analyzing the Adult Dataset

7-15-2022

Teammate Names: Matthew Long, Bin Gui, Lance

### Data Processing

First we obtained the dataset from the UCI Machine Learning Repository. We then uploaded this set into a dataframe. The 50k category was one-hot encoded so that it could store numerical values. Finally, we separated the dataset into a training and testing set.

### Model

We implemented both a decision tree model and a logistic regression neural network for this data set since it was a classification problem.

### Evaluation

For our evaluation, we used the metrics of entropy and accuracy. To streamline our results, we will focus on accuracy as the basis of our evaluation.

### Test

After creating, training, and testing the models, the target value was classified at 79% accuracy with the decision tree and logistic regression performing at the same levels.

# Roadmap | Executive Diagram

Teammate Names: Matthew Long, Bin Gui, Lance

# Dataset Overview

- The dataset 48842 instances and 14 attributes relating to an adults.
- Attributes include age, education, occupation, education number, race, etc.
- Our target attribute is whether the income is over or under $50k.
- We will be testing different features to reach an accurate prediction.

Teammate Names: Matthew Long, Bin Gui, Lance

# Data Display

| | age | workclass | fnlwgt | education | education-num | marital-status | occupation | relationship | race | sex | capital-gain | capital-loss | hours-per-week | native-country | 50k |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **32556** | 27 | Private | 257302 | Assoc-acdm | 12 | Married-civ-spouse | Tech-support | Wife | White | Female | 0 | 0 | 38 | United-States | <=50K |
| **32557** | 40 | Private | 154374 | HS-grad | 9 | Married-civ-spouse | Machine-op-inspct | Husband | White | Male | 0 | 0 | 40 | United-States | >50K |
| **32558** | 58 | Private | 151910 | HS-grad | 9 | Widowed | Adm-clerical | Unmarried | White | Female | 0 | 0 | 40 | United-States | <=50K |
| **32559** | 22 | Private | 201490 | HS-grad | 9 | Never-married | Adm-clerical | Own-child | White | Male | 0 | 0 | 20 | United-States | <=50K |
| **32560** | 52 | Self-emp-inc | 287927 | HS-grad | 9 | Married-civ-spouse | Exec-managerial | Wife | White | Female | 15024 | 0 | 40 | United-States | >50K |

**Teammate Names: Matthew Long, Bin Gui, Lance**

# Model Construction

Our first model was a decision tree. We tried to implement different features, but we found that education-num provided the highest accuracy. Other features, such as age and work hours, created much more expansive trees with lower accuracies.

**Fig 1: Decision tree for work hours/week**

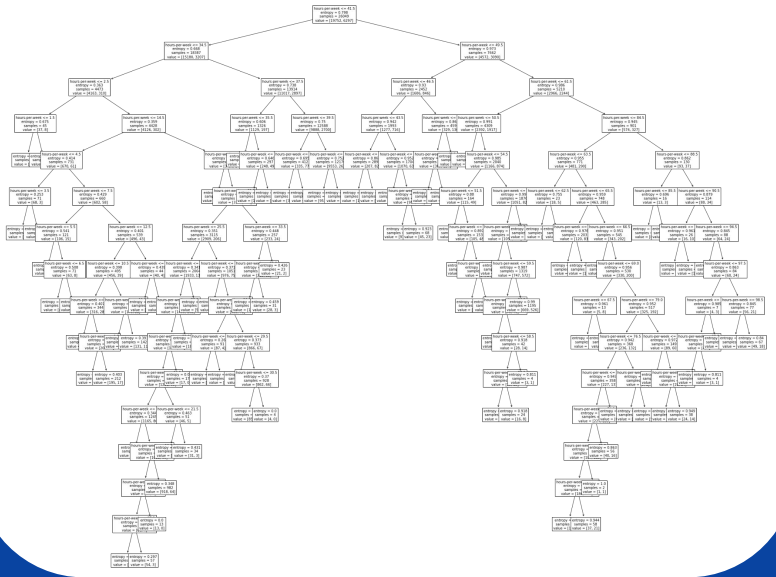
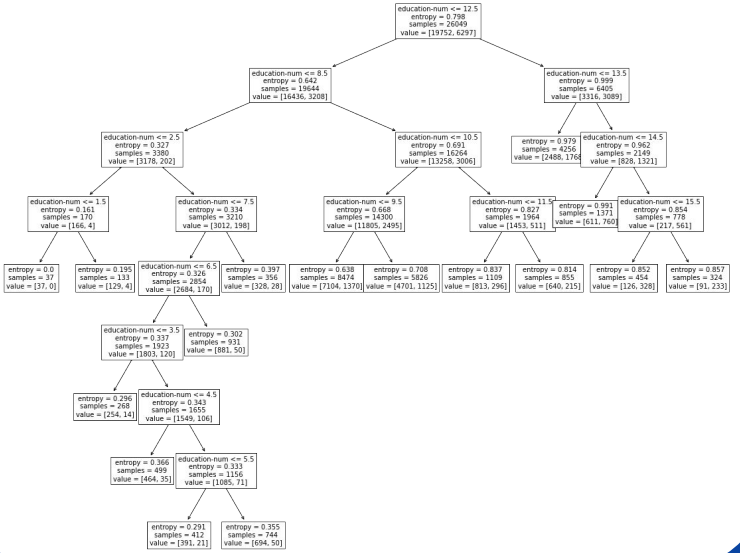


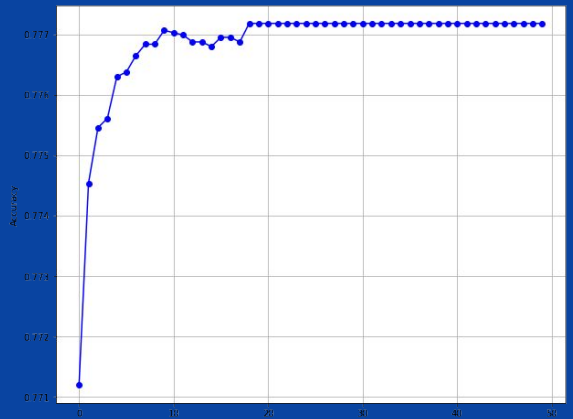
**Fig 2: Final Decision tree for education num**

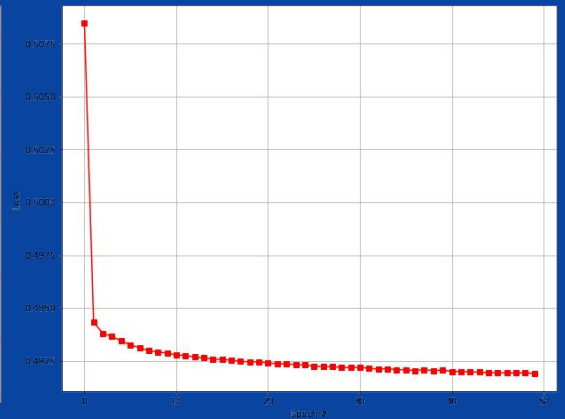


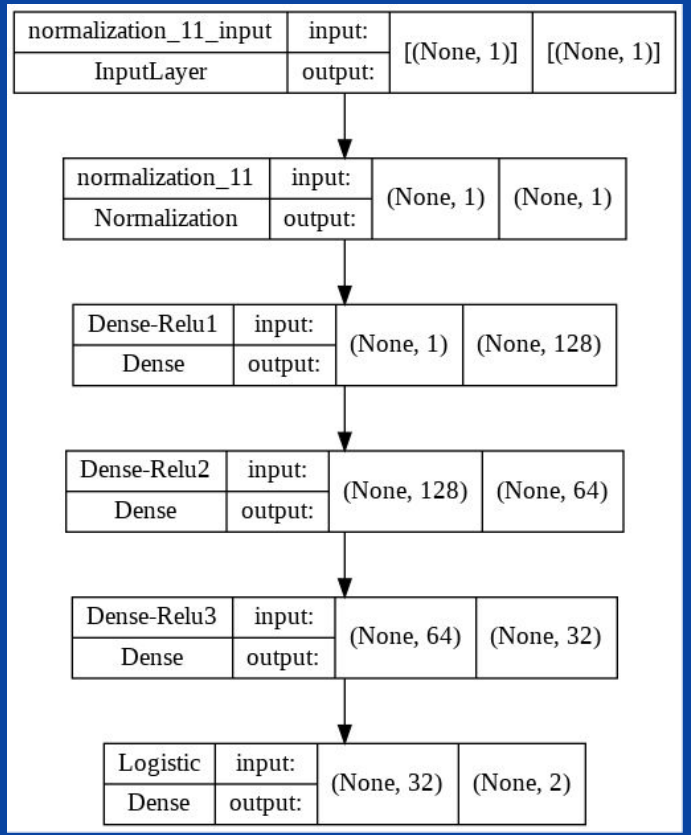
Teammate Names: Matthew Long, Bin Gui, Lance

# Model Construction

Our second model was a logistic regression neural network. Similar to the decision tree, we tested different features. Since this model was a neural network, we also tried out different layers.







Teammate Names: Matthew Long, Bin Gui, Lance

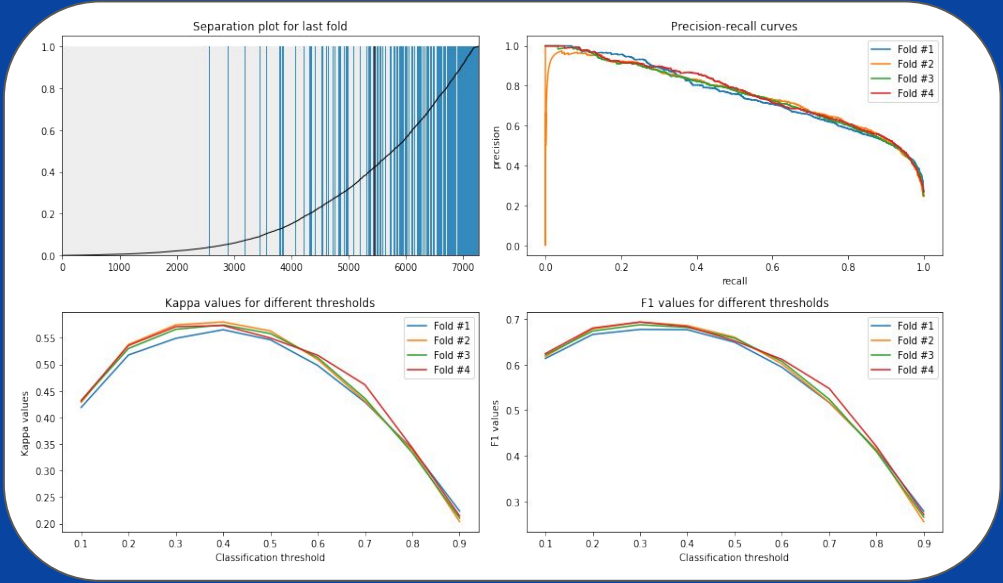
# Performance

| | Decision Tree | | | Logistic Regression | | |
|---|---|---|---|---|---|---|
| | Feature 1 | Feature 2 | Feature 3 | Feature 1 | Feature 2 | Feature 3 |
| Training | 0.76 | 0.76 | 0.78 | 0.7587 | 0.7587 | 0.7772 |
| Validation | N/A | | | 0.7566 | 0.7566 | 0.7770 |
| Test | 0.76 | 0.76 | 0.79 | 0.7636 | 0.7627 | 0.7902 |

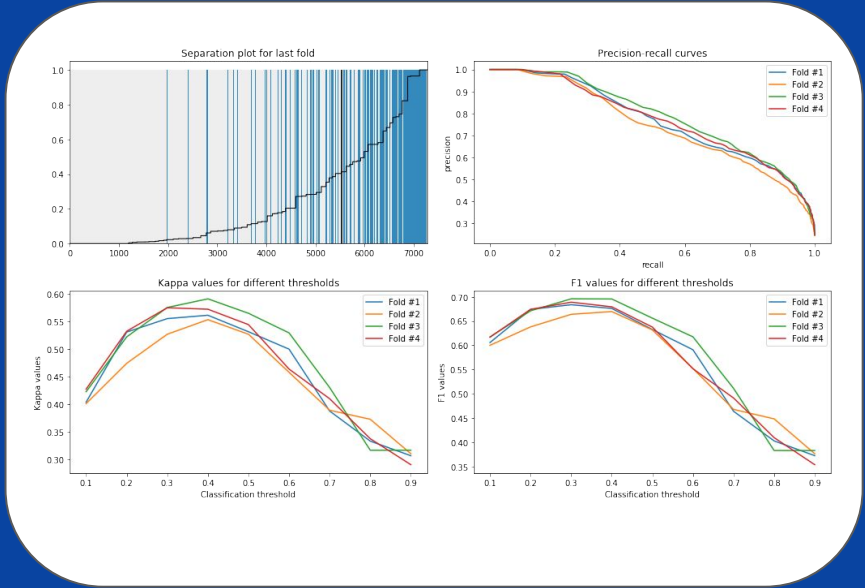
Feature 1: Hours per week
Feature 2: Age
Feature 3: Education-num

# Additional Performances

Logistic Regression: 0.841




Decision Tree: 0.843




https://github.com/itdxer/adult-dataset-analysis/blob/master/Classification.ipynb

# Conclusions

- Both our models tied with a 79% accuracy rate for the test set.
- Feature 3 proved to be the most accurate in determining the target.
- Feature 1 and 2 performed nearly identically across both models.
- The outside source further demonstrated that both models produced similar results in terms of accuracy.
- From this project, we have learned that data, especially large sets, often form patterns when analyzed through machine learning models. The similarity between performances demonstrates how creating models can be unpredictable and sometimes no singular model is the “correct” choice.