Salary Prediction: Logistic Regression

Learning Objectives:

- *Data cleaning for multi-class classification*
- *Feature selection algorithms and regularization*
- *Logistic regression*
- *Model selection and bias-variance trade-off*

Samuel Atkins - November 2020

Exploratory Data Analysis + Feature Selection

Original Dataset Size: *12497 x 248*

Question Categories:

1. Select all that apply
   - Transformed into one-hot columns
2. Multiple choice/categorical
   - Situationally turned into one-hot columns or ordinally encoded columns
3. Numerical
   - Scaled or left alone

Columns with significant null values were removed; null entries were replaced with column means/modes

Column Dependence Tests:

a. One-hot encoded multiple-choice columns tested using chi^2 test for independence

b. Ordinally encoded and numerical columns tested using ANOVA/F-test

c. Select all that apply binary columns tested using chi^2 squared test for independence

- Features with a p-value greater than 0.05 were removed

RFE was then applied to extract the top 150 features

Final Dataset Size: *12497 x 150*

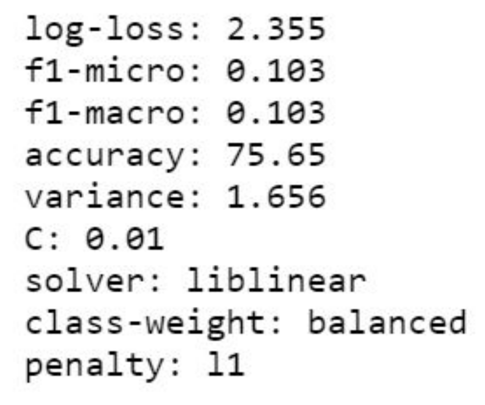
Logistic Regression

- Accuracy and variance selected as primary performance metrics
  - F1-micro, F1-macro,and log-loss also included for performance visualization
- Grid-search performed to tune the regularization constant, the solver, the regularization penalty type, and the class weight setting

Grid Search Process:

Iterate through all possible models to find the model that maximizes the accuracy (minimizes the bias) and simultaneously minimizes the variance through 10-fold cross-validation on the training dataset

Optimal Model:



```
log-loss: 2.355
f1-micro: 0.103
f1-macro: 0.103
accuracy: 75.65
variance: 1.656
C: 0.01
solver: liblinear
class-weight: balanced
penalty: l1
```

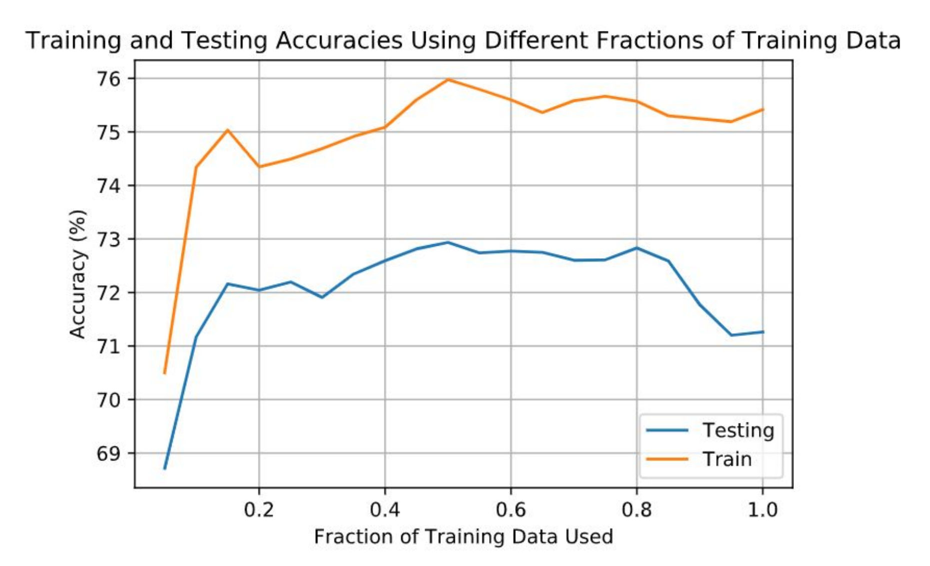
Bias-Variance Trade-off

Bias-Variance Trade-off Investigation:

- Trained the optimal model using various percentages of the training dataset
- Computed the accuracy of the model on the entire training dataset and the entire testing dataset using each trained model
- Plotted the training-testing curve

Observation: *Model is Overfitting*

- Could implement early stopping to improve generalizability



Model Testing + Discussion

%

Testing Result:

Model Achieved a Testing Accuracy of 71.287% on the Holdout Dataset

Discussion:

- Models with less features typically performed significantly worse
- Early stopping would improve the generalizability of the model
- Learning about how the model might be used would incentivize a more fine-tuned performance metric
- Perhaps a more complex or learned model could capture the relationships between the input data and target variables more effectively