Hackathon

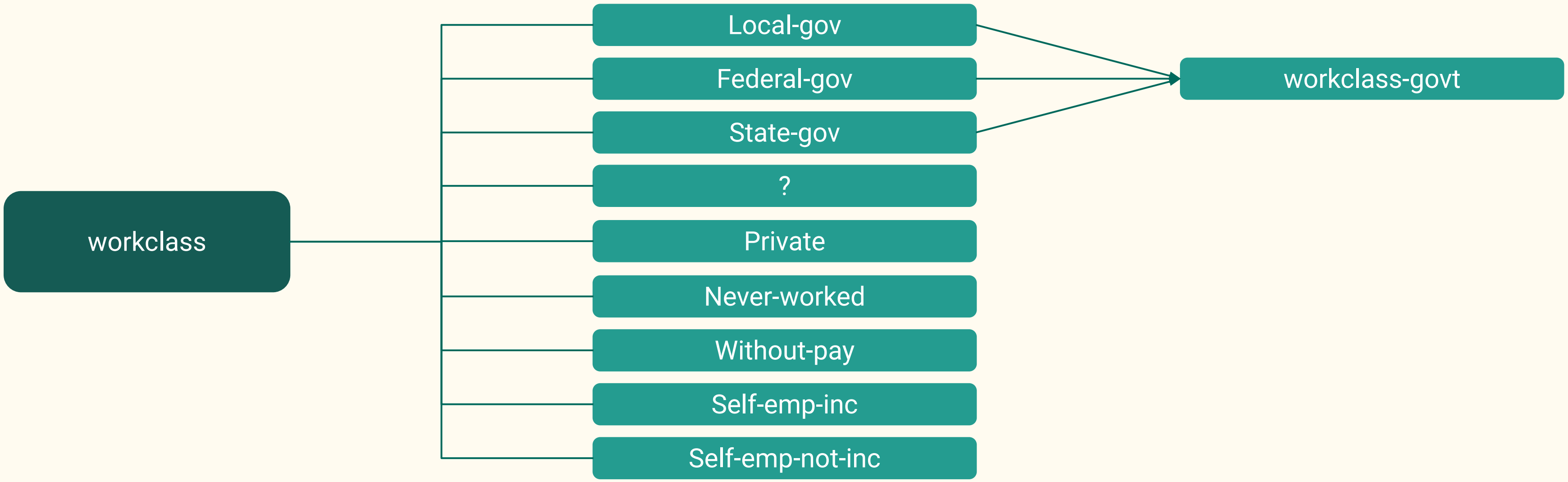
Tiffany Houston, Mack McGowen, Jared Delora-Ellefson

Overview

Problems Statement: Given the data at hand, when constrained to 20 features, how well can we predict if a person makes greater than $50k/yr.

- Feature Engineering
- Visualizations
- Modeling Process
- Metrics Summary

Feature Engineering


workclass

Local-gov

Federal-gov

State-gov

?

Private

Never-worked

Without-pay

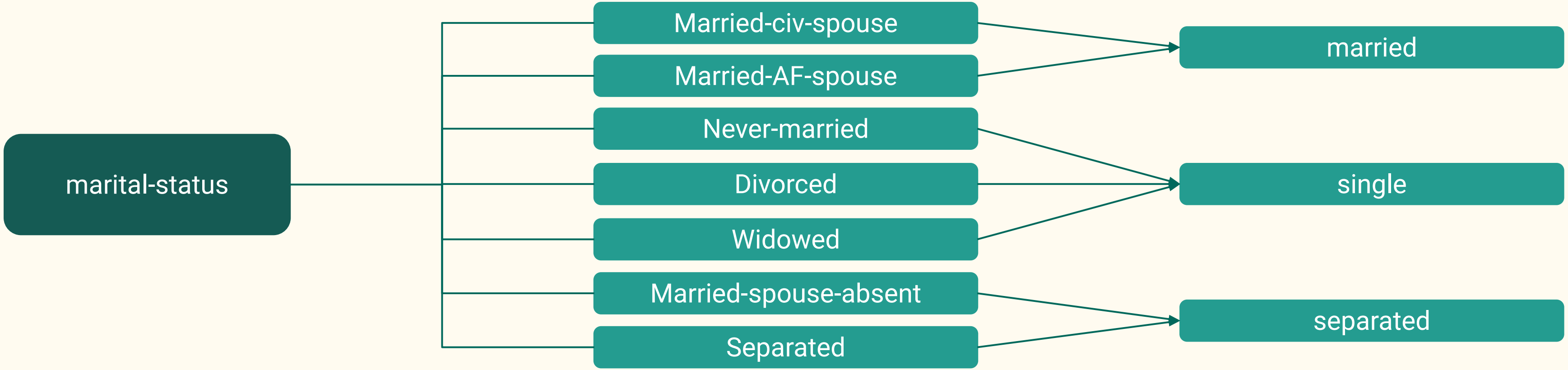
Self-emp-inc

Self-emp-not-inc

workclass-govt

Government workers are grouped to save feature space due to constraints

Feature Engineering (cont)



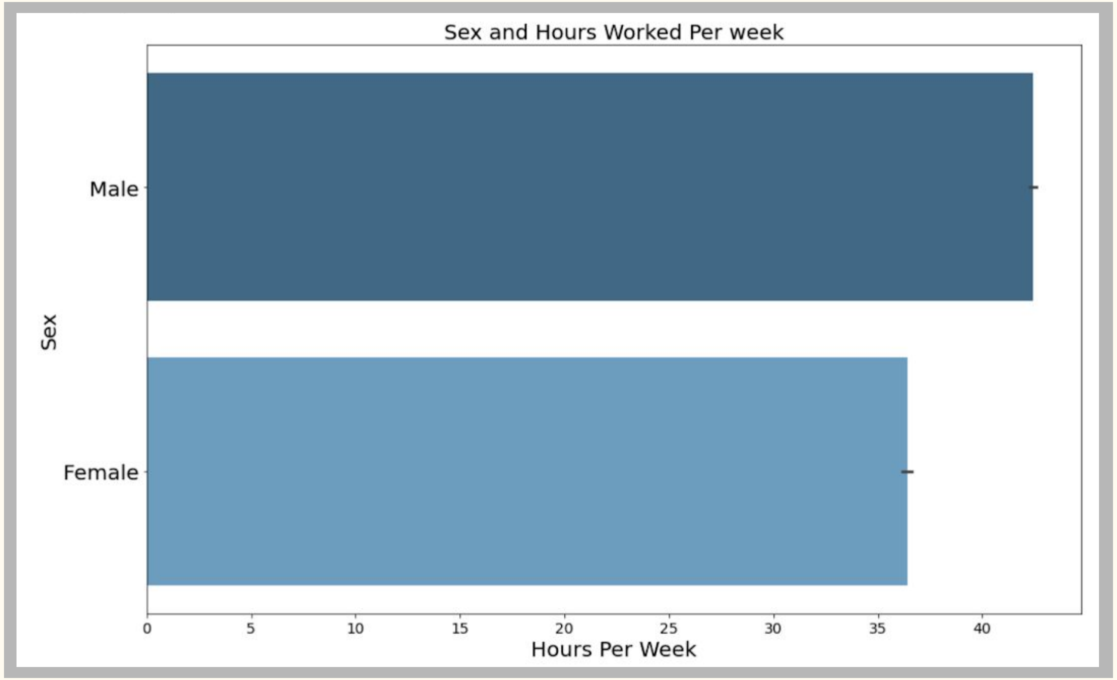
Marital statuses are grouped into 3 features due to constraints

Hours Worked Per Week by Sex:

Male ~ 43 hrs/wk

vs

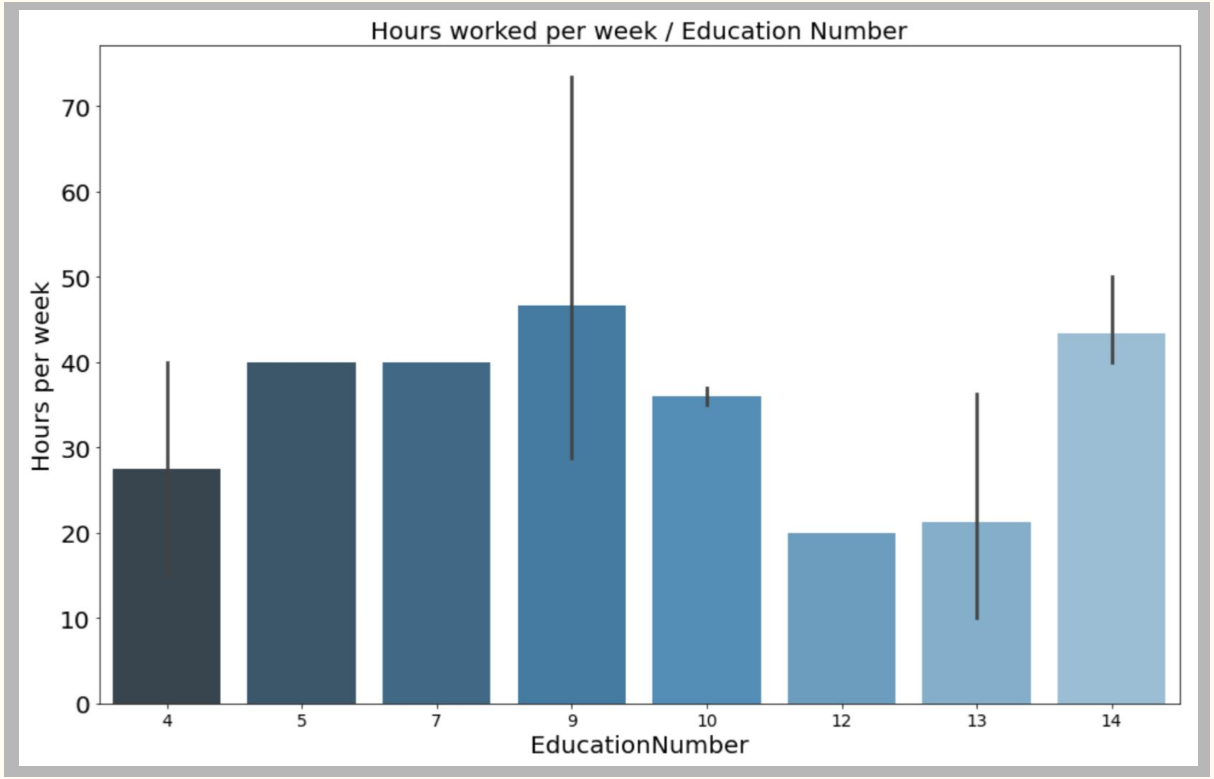
Female ~ 36 hrs/wk



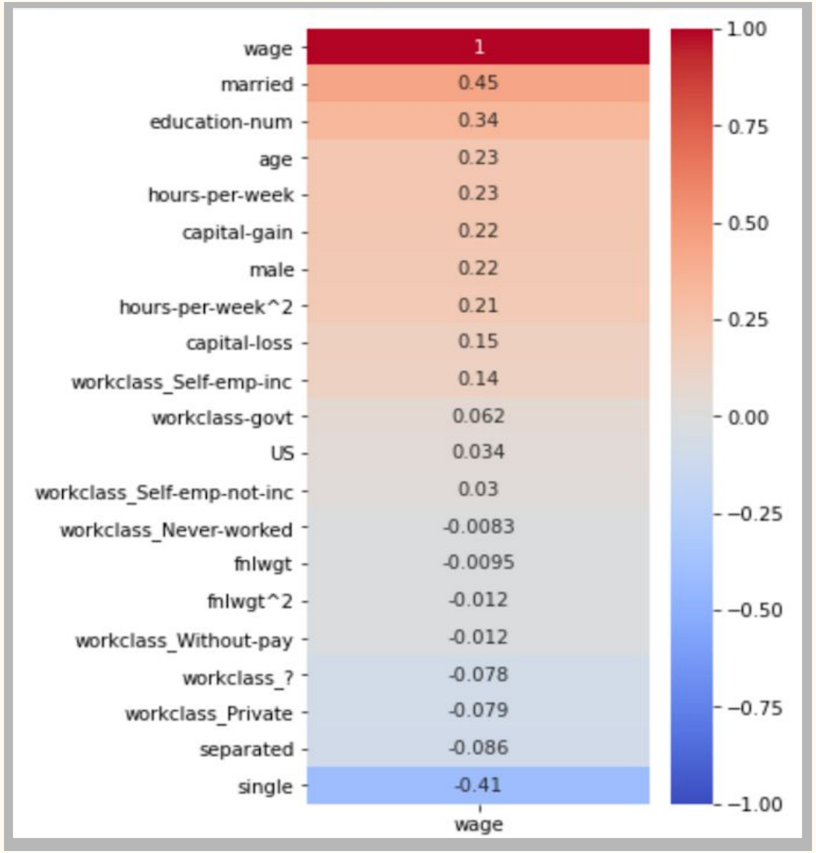
Hours Worked per Week by Education Number:

Top 3 numbers working most hours:

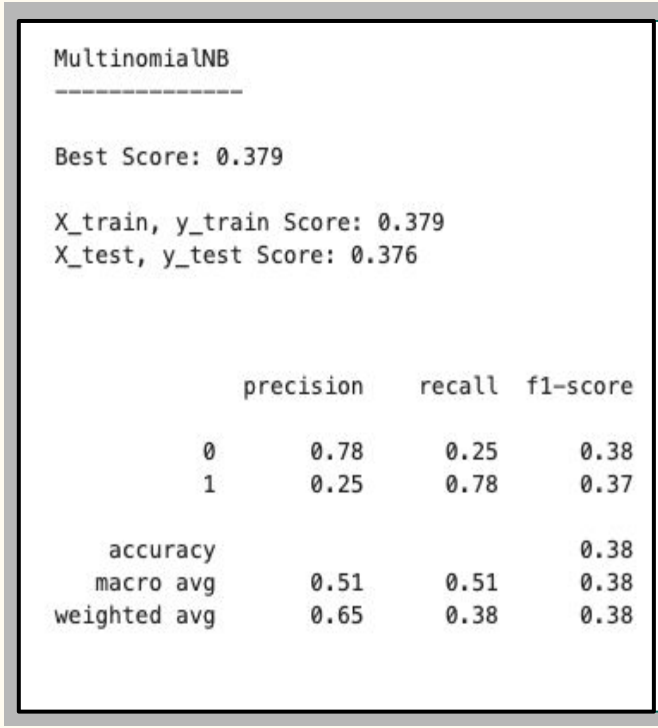
- 9 (47 hrs/wk)
- 14 (45 hrs/wk)
- 5/7 (40 hrs/wk)



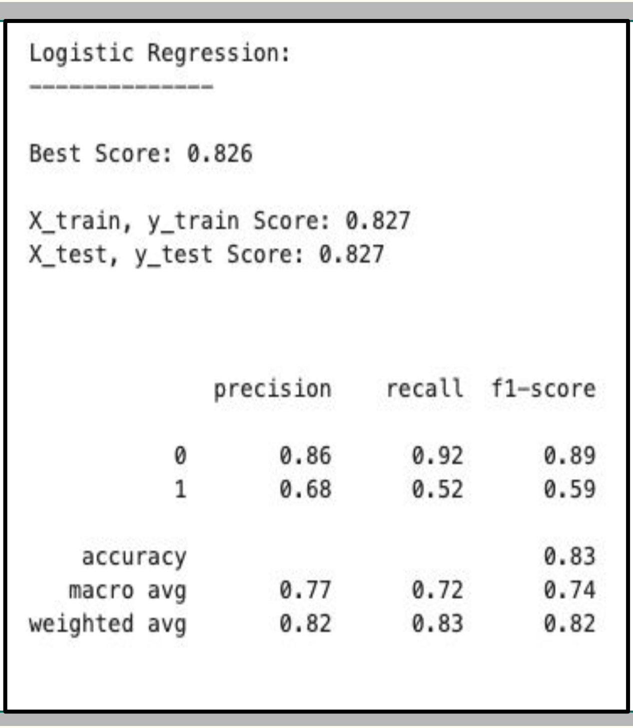
How Factors Correlate to Wage



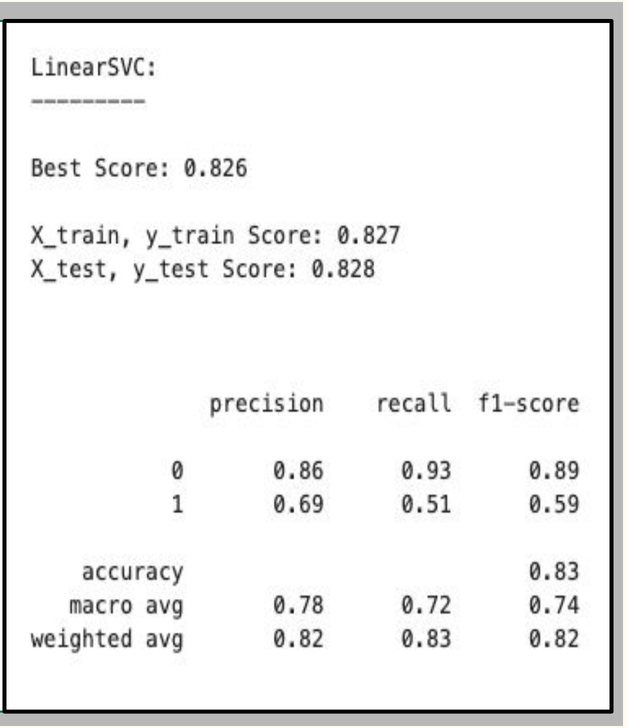
Naive Bayes, Logistic, and SVC Classification



```
MultinomialNB
--------------

Best Score: 0.379

X_train, y_train Score: 0.379
X_test, y_test Score: 0.376


             precision    recall  f1-score

           0       0.78      0.25      0.38
           1       0.25      0.78      0.37

    accuracy                           0.38
   macro avg       0.51      0.51      0.38
weighted avg       0.65      0.38      0.38
```



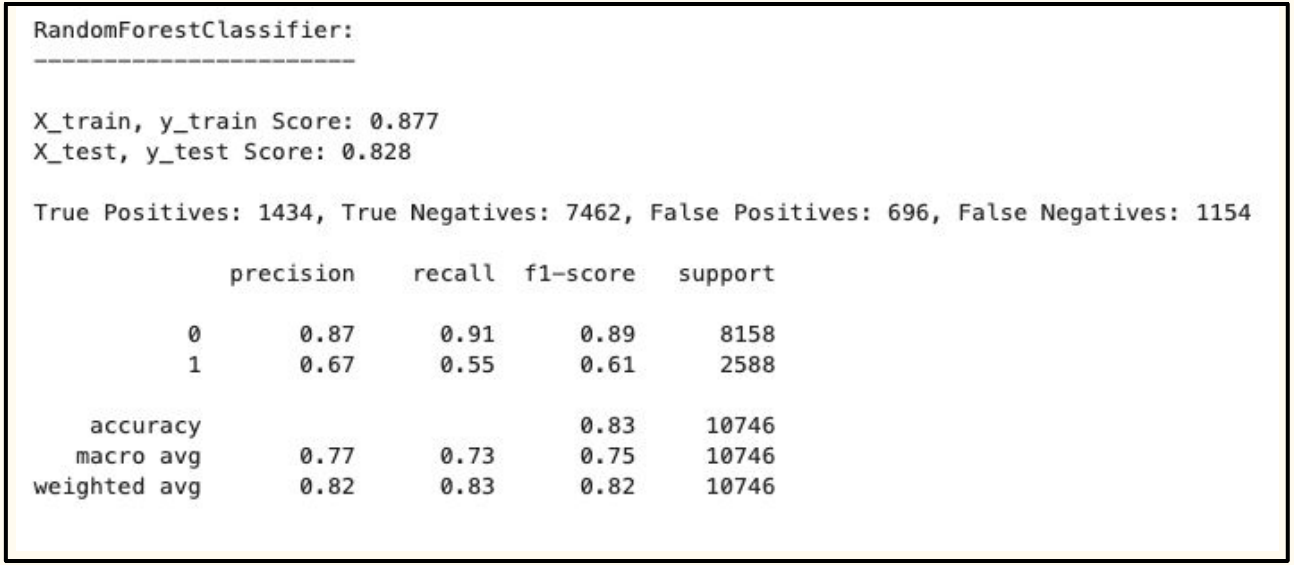
```
Logistic Regression:
--------------

Best Score: 0.826

X_train, y_train Score: 0.827
X_test, y_test Score: 0.827


             precision    recall  f1-score

           0       0.86      0.92      0.89
           1       0.68      0.52      0.59

    accuracy                           0.83
   macro avg       0.77      0.72      0.74
weighted avg       0.82      0.83      0.82
```



```
LinearSVC:
---------

Best Score: 0.826

X_train, y_train Score: 0.827
X_test, y_test Score: 0.828


             precision    recall  f1-score

           0       0.86      0.93      0.89
           1       0.69      0.51      0.59

    accuracy                           0.83
   macro avg       0.78      0.72      0.74
weighted avg       0.82      0.83      0.82
```

Random Forest Classifier

- This model performed the best.
- There is an imbalance in the target variable (0.76% of the target variables were ≥ 50K).
- Due to this imbalance, predicting ≥ 50K has a lower precision than $<$ 50K.



```
RandomForestClassifier:
-----------------------

X_train, y_train Score: 0.877
X_test, y_test Score: 0.828

True Positives: 1434, True Negatives: 7462, False Positives: 696, False Negatives: 1154

              precision    recall  f1-score   support

           0       0.87      0.91      0.89      8158
           1       0.67      0.55      0.61      2588

    accuracy                           0.83     10746
   macro avg       0.77      0.73      0.75     10746
weighted avg       0.82      0.83      0.82     10746
```

Model Metrics Summary

Model	Target - Income	Precision	Recall	F1 Score	Accuracy
Multinomial Naive Bayes	<= 50K	0.78	0.25	0.38	0.38
	> 50K	0.25	0.78	0.37	
Logistic Regression	<= 50K	0.86	0.92	0.89	0.83
	> 50K	0.68	0.52	0.59	
Support Vector Classifier	<= 50K	0.86	0.91	0.89	0.83
	> 50K	0.69	0.51	0.59	
Random Forest Classifier	<= 50K	0.87	0.91	0.89	0.83
	> 50K	0.67	0.55	0.61	