Project 4 Hackathon:
Feature Constraint

Stephen Strawbridge
Aziz Maredia
Mike Winder
Max Bosse

Overview

1. Project Scope + Constraint
2. Feature Engineering
3. Models and Model Performance
4. Next Steps and Conclusions

Project Task and Constraint

Goal: Predict if a person's income is in excess of $50,000 given profile information.

Constraint: Limited to maximum of 20 features in model

- Cannot Dummify all categorical features

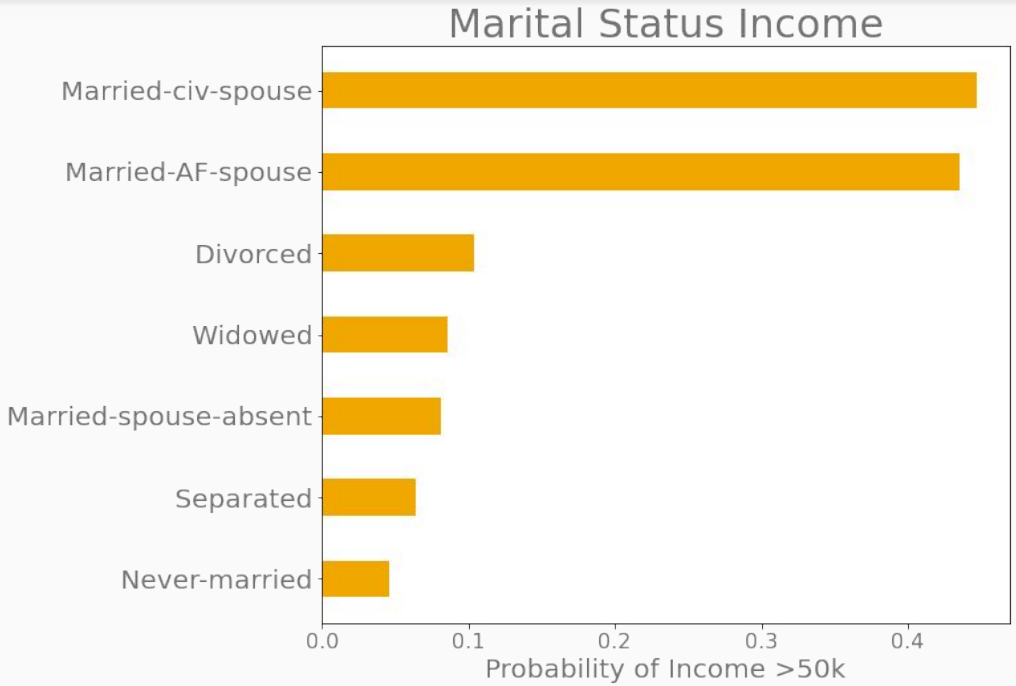
Profile Information Provided

- Age
- Workclass
- Education Level
- Marital Status
- Occupation
- Sex
- Native Country
- Capital Gain/Loss

Feature Engineering - Dealing with Categorical Variables

To Conserve number of Features:

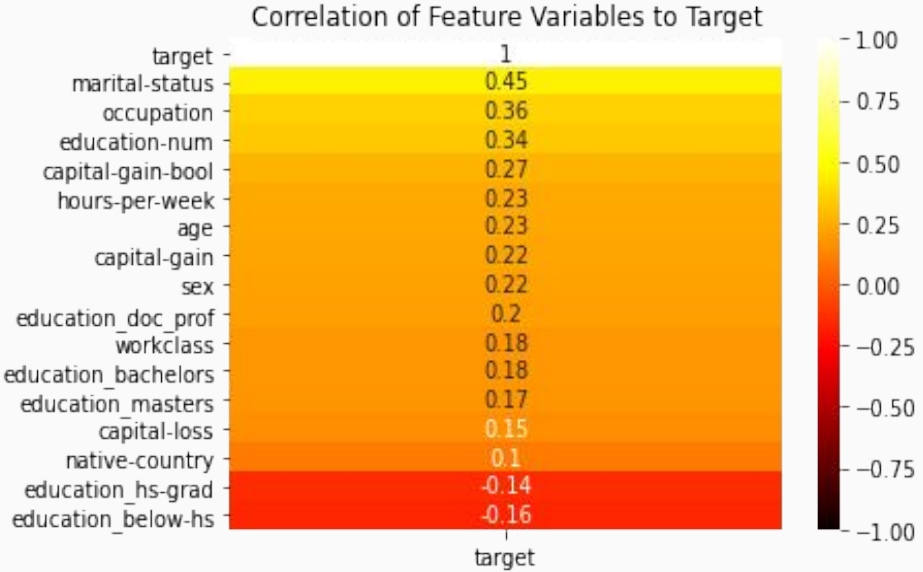
Instead of using dummies, Impute probabilities



Experimenting with Columns

Example of Dummy work-around given constraint:

Below HS - Preschool - 1st-4th - 5th-6th - 7th-8th - 9th - 10th - 11th	**HS Grad** - 12th - HS-grad	**Assoc_Some_College** - Assoc-acm - Assoc-voc - some-college
Bachelors -bachelors	**Masters** -masters	**Doc_prof** -Doctorate - Prof-school



Models and Model Performance

Model	Model Score
AdaBoost	0.8604
Random Forest	0.8589
Bagging Classifier	0.8498
Logistic Regression	0.8056
Decision Tree	0.8050
Support Vector Machine	0.8045
K Nearest Neighbors	0.7893
Baseline	*0.7592*

Conclusions and Next Steps

Best Scoring Model → AdaBoost

- Train: ~87%
- Test: ~86%

Best Fit Model → Logistic Regression

- Train: ~83%
- Test: ~83%

Top predictors in Logistic Regression:

- Marital-Status
- Occupation

Next Steps -

- Try testing numeric columns as dumified categorical variables
- Dummify or binarize other variables rather than imputing probabilities
- Engineer and test additional columns
  - Ex. Interaction Terms, Poly Features

Thank you!

Any questions?