Hackathon Good Fast Cheap

Kaytlynn Skibo
Chris Landschoot
Nicholas Nguyen
Ayako Homma

March 7, 2022

Problem statement

The goal is to create a best performing model on a ‘census income’ data and predict whether a person’s income exceeds $50,000 a year, given certain profile information.

Challenge: Cheap Training Data | Smaller dataset than others (20%)

How success is measured: Accuracy

Key steps in the research process

01 Data Import & Cleaning

02 EDA

03 Modeling & Evaluation

04 Conclusion & Recommendation

Data import & cleaning

Data Import

Cheap_train_sample (6513 x 14)

Data Cleaning

Map '?' to nan or replace with mode

- 'Native-country' > mode (United States)
- 'Workclass' > mode ('Private')
- 'Occupation' > nan

Convert to dummy variables

- 'Martial status', 'occupation', 'relationship', 'native country', 'workclass'

Resampling

For small datasets

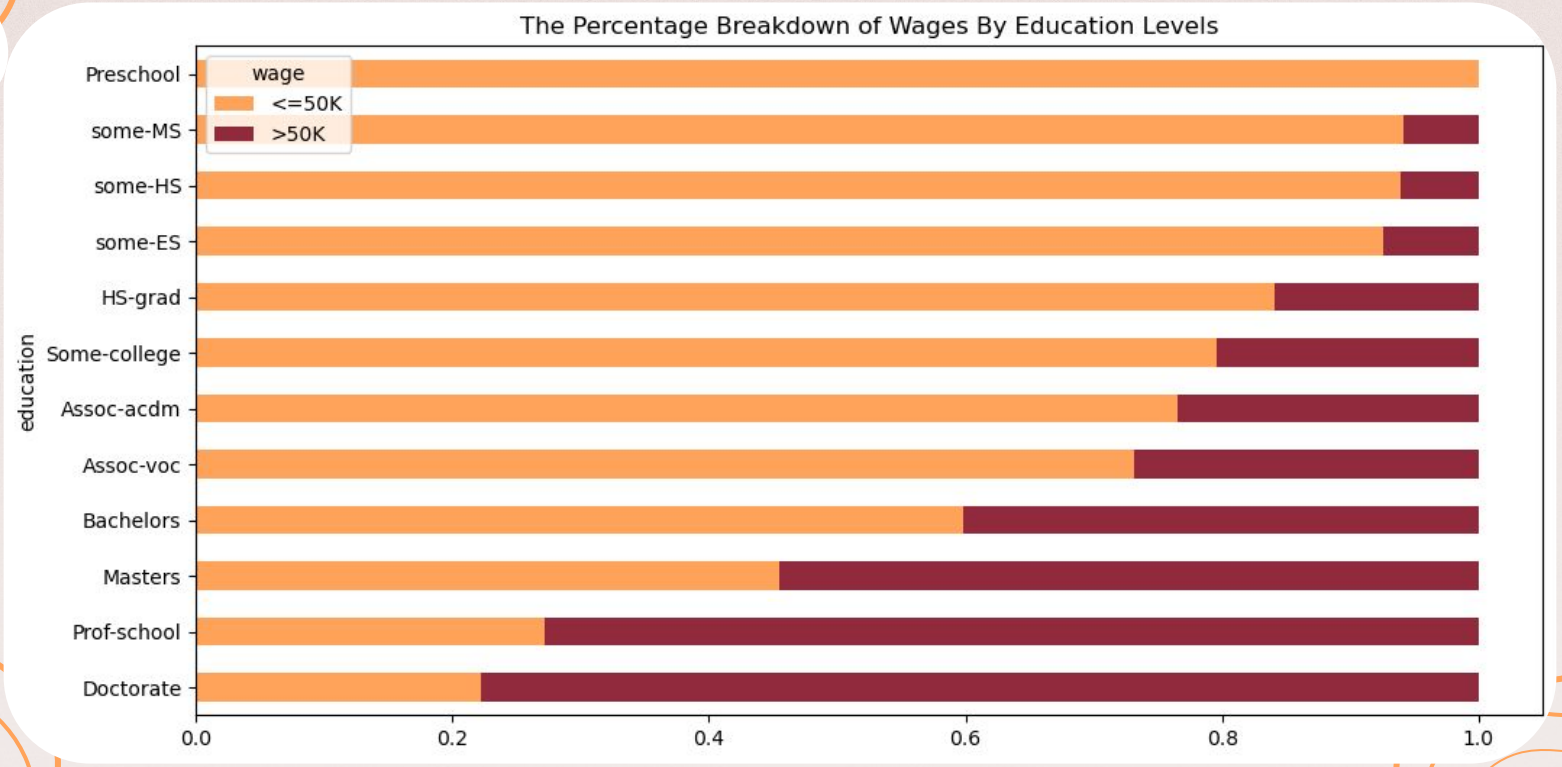
Fixing:

- Larger dataset to evaluate
- Balancing skew
  - .24 > .49
- Improve accuracy

6513 samples

39713 samples

Doctorate & prof-school graduates show higher percentage of incomes above $50,000 per year

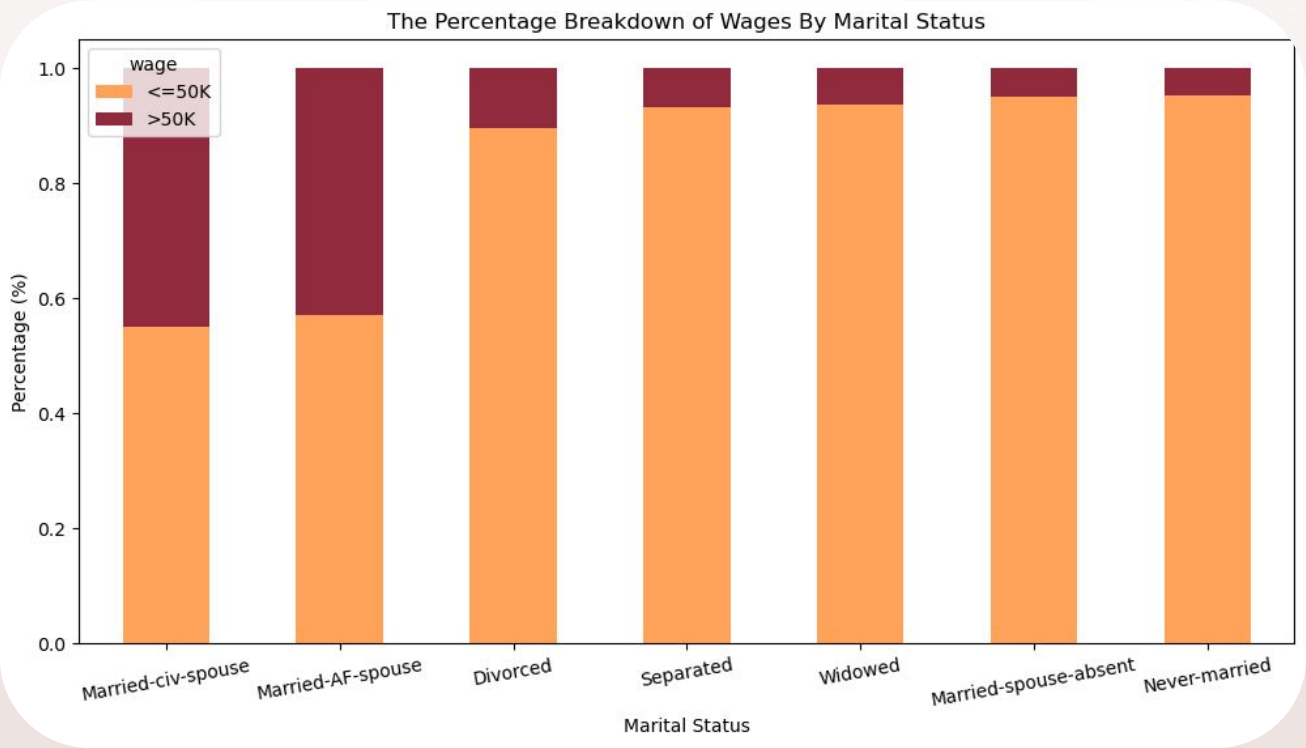


Source: US census income data

Exec-managerial & prof-specialty occupations show higher percentage of wages above $50,000 per year

Source: US census income data

Married individuals with spouses show higher percentage of wages above $50,000 per year



Source: US census income data

Data modeling & evaluation

Model Type	Train Accuracy	Test Accuracy	Specificity	Precision	Sensitivity
Logistic Regression	0.823	0.814	0.796	0.797	0.833
KNN	1.00	0.991	0.984	0.983	0.998
SVM	0.875	0.866	0.816	0.827	0.918

Baseline: 0.508

Conclusion & recommendations

Conclusions

- Training data was bootstrapped
- KNN outperformed other models
- The model is overfit

Recommendations

- Test other model types
- Collect more data
  - Increase existing set
  - New categories
- Engineer new features

Thank you!

Any questions?