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

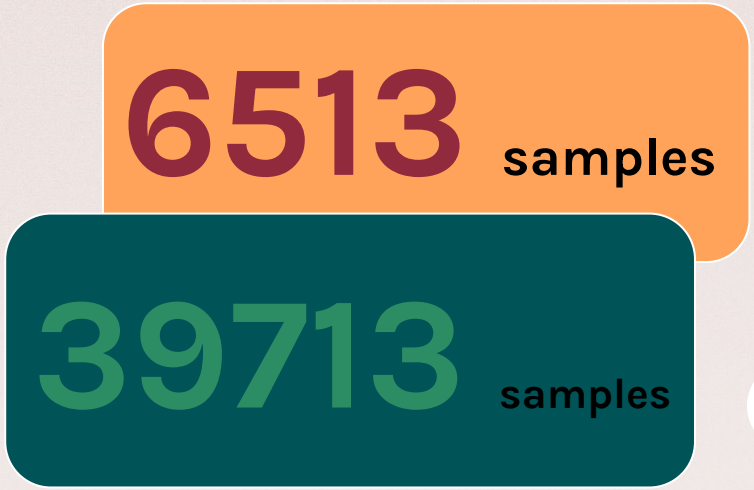
Fixing:

Larger dataset to evaluate

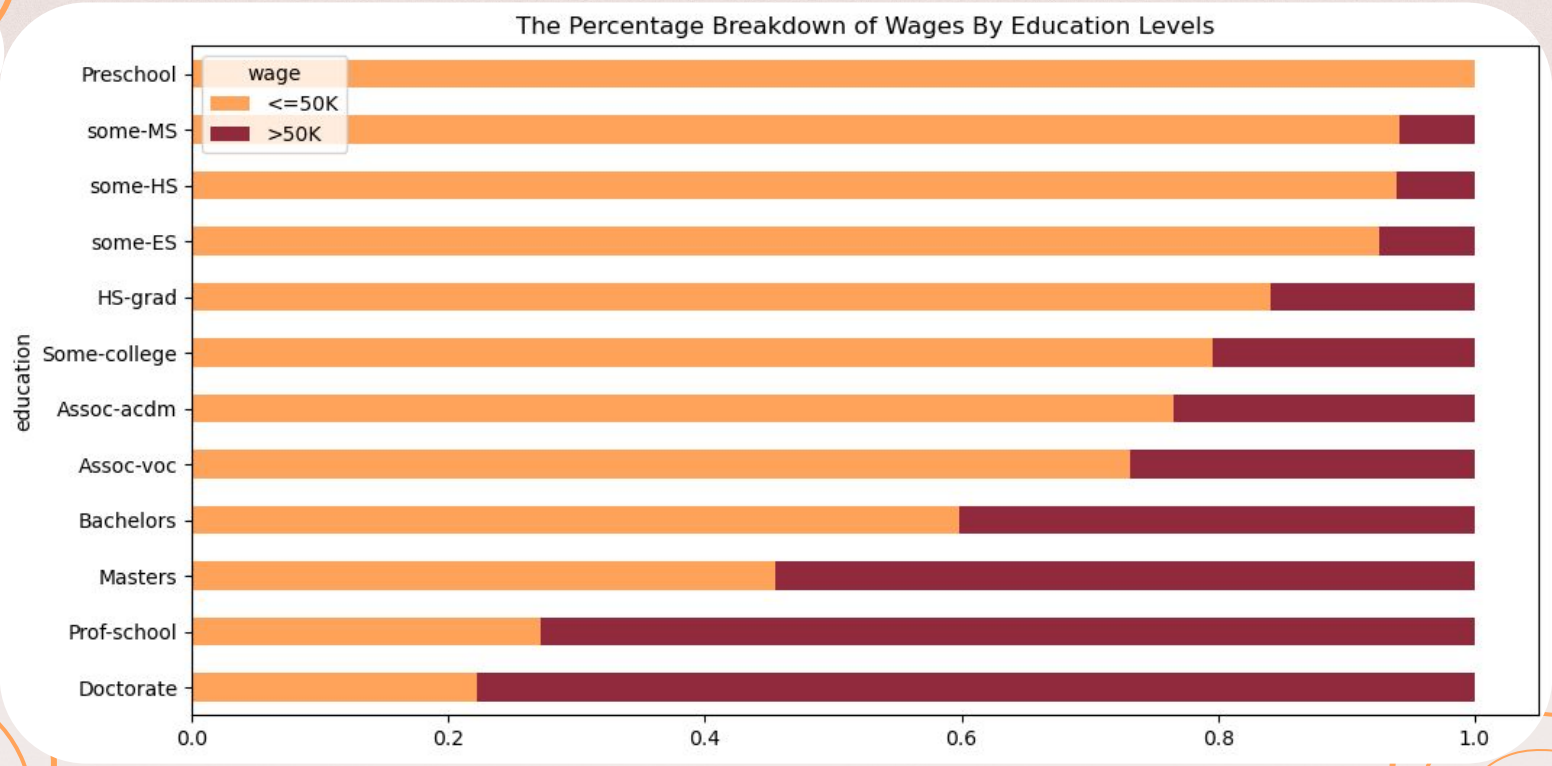
Balancing skew

.24 > .49

Improve accuracy

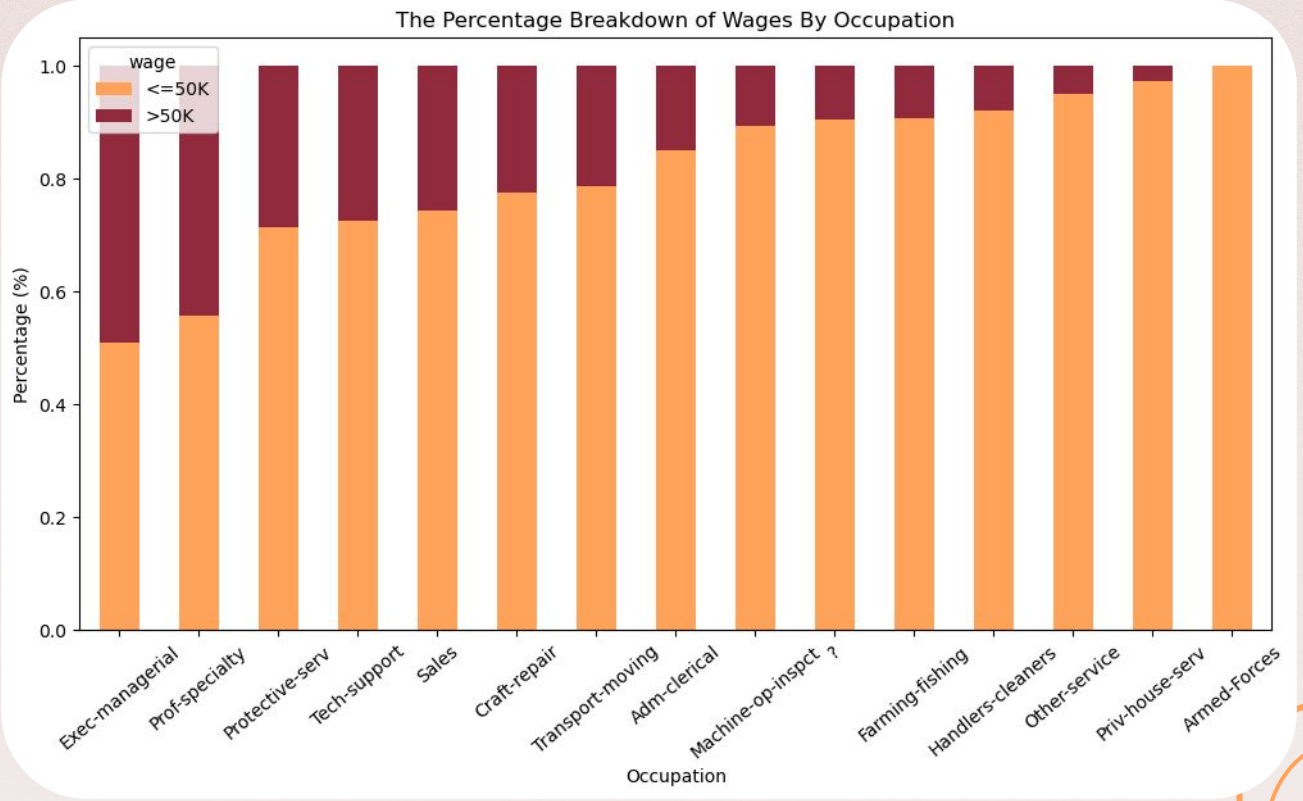


EDA

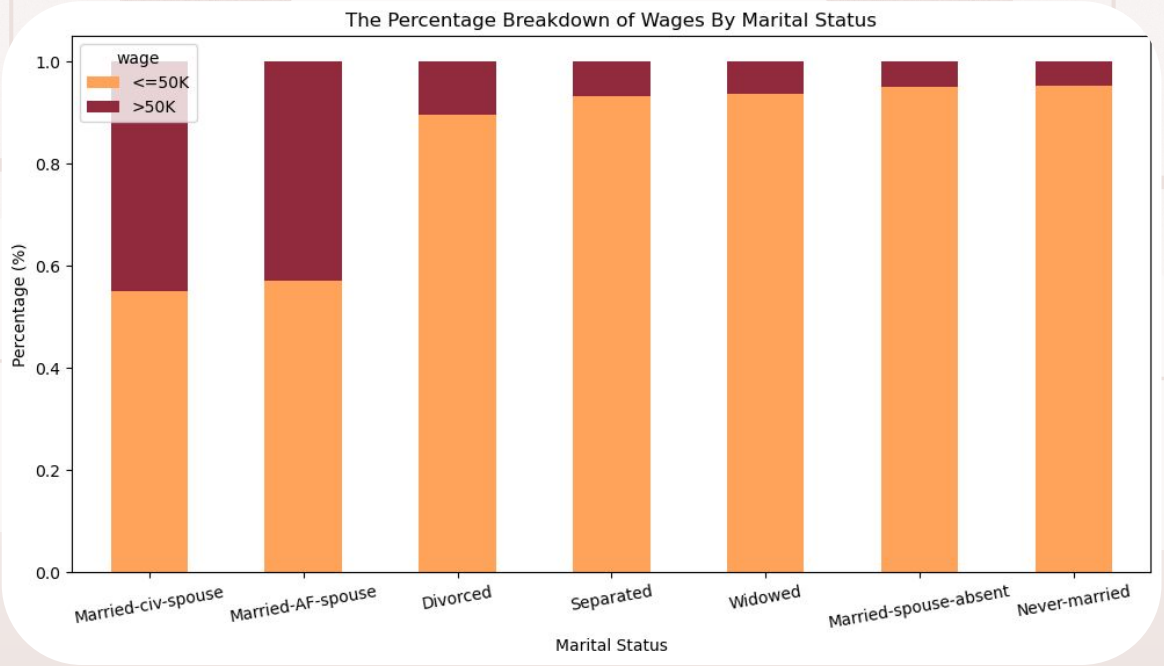


Source: ‘census income’ data

EDA


The Percentage Breakdown of Wages By Occupation
wage
<=50K
>50K
Percentage (%)
1.0
0.8
0.6
0.4
0.2
0.0
Exec-managerial
Prof-specialty
Protective-serv
Tech-support
Sales
Craft-repair
Transport-moving
Adm-clerical
Machine-op-inspct
?
Farming-fishing
Handlers-cleaners
Other-service
Priv-house-serv
Armed-Forces
Occupation

EDA


The Percentage Breakdown of Wages By Marital Status
wage
<=50K
>50K
Percentage (%)
1.0
0.8
0.6
0.4
0.2
0.0
Married-civ-spouse
Married-AF-spouse
Divorced
Separated
Widowed
Married-spouse-absent
Never-married
Marital Status

Modeling

Baseline: 0.508

Model Type	Train Accuracy	Test Accuracy	Specificity	Accuracy	Precision	Sensitivity
Logistic Regression	0.824	0.817	0.799	0.817	0.799	0.857
KNN	1.00	0.992	0.986	0.992	0.985	0.998
SVM	0.870	0.861	0.806	0.861	0.820	0.522

Conclusion & Recommendation

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?