پروژه داده کاوی

سعید قاسمی
رسول کامکار

فهرست مطالب

مقدمه

بررسی مسئله

- پیشبینی سطح درآمد افراد
- بیشتر یا کمتر از 50,000$
- مسئله classification

داده‌های جمع‌آوری شده[1]

- Age (continuous integer)
- Workclass (categorical)
- Fnlwgt [2] (continuous integer)
- Education (categorical)
- Education-num (continuous integer)
- Martial-status (categorical)
- Occupation (categorical)

داده‌های جمع‌آوری شده (ادامه)

- Relationship (categorical)
- Race (categorical)
- Gender (categorical)
- Capital-gain (continuous integer)
- Capital-loss (continuous integer)
- Hours-per-week (continuous integer)
- Native-country (categorical)

خلاصه‌ای از روند کلی حل مسئله

- بررسی داده در دست
- پیش پردازش داده‌ها
- انتخاب و train مدل‌ها
- انتخاب مدل با عملکرد بهتر

نحوه ارزیابی

- انتخاب کلاس با درصد بیشتر به عنوان baseline
- بررسی Accuracy نسبت به baseline
- بررسی معیارهای F1 score و precision

فهرست مطالب

آزمایشات

بررسی داده

بررسی ویژگی‌های داده

- ۱۵ ستون (۱۴ feature و یک متغیر هدف)
- ۴۳۹۵۷ عدد رکورد موجود
- تا ۴۱ مقدار یکتا در خصیصه‌های categorical

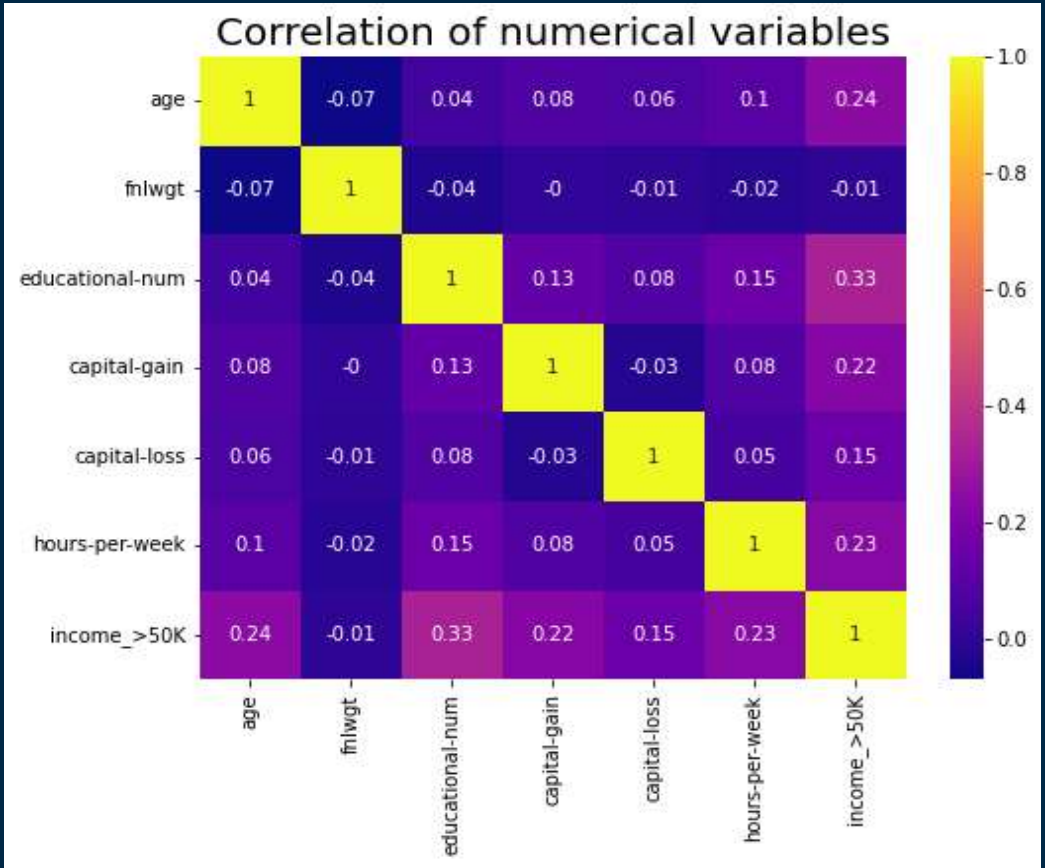
بررسی ویژگی‌های داده

- متغیر هدف income_>50K از نوع binary
- ۷۵ درصد کلاس 0 (baseline)
- ۲۵ درصد کلاس 1

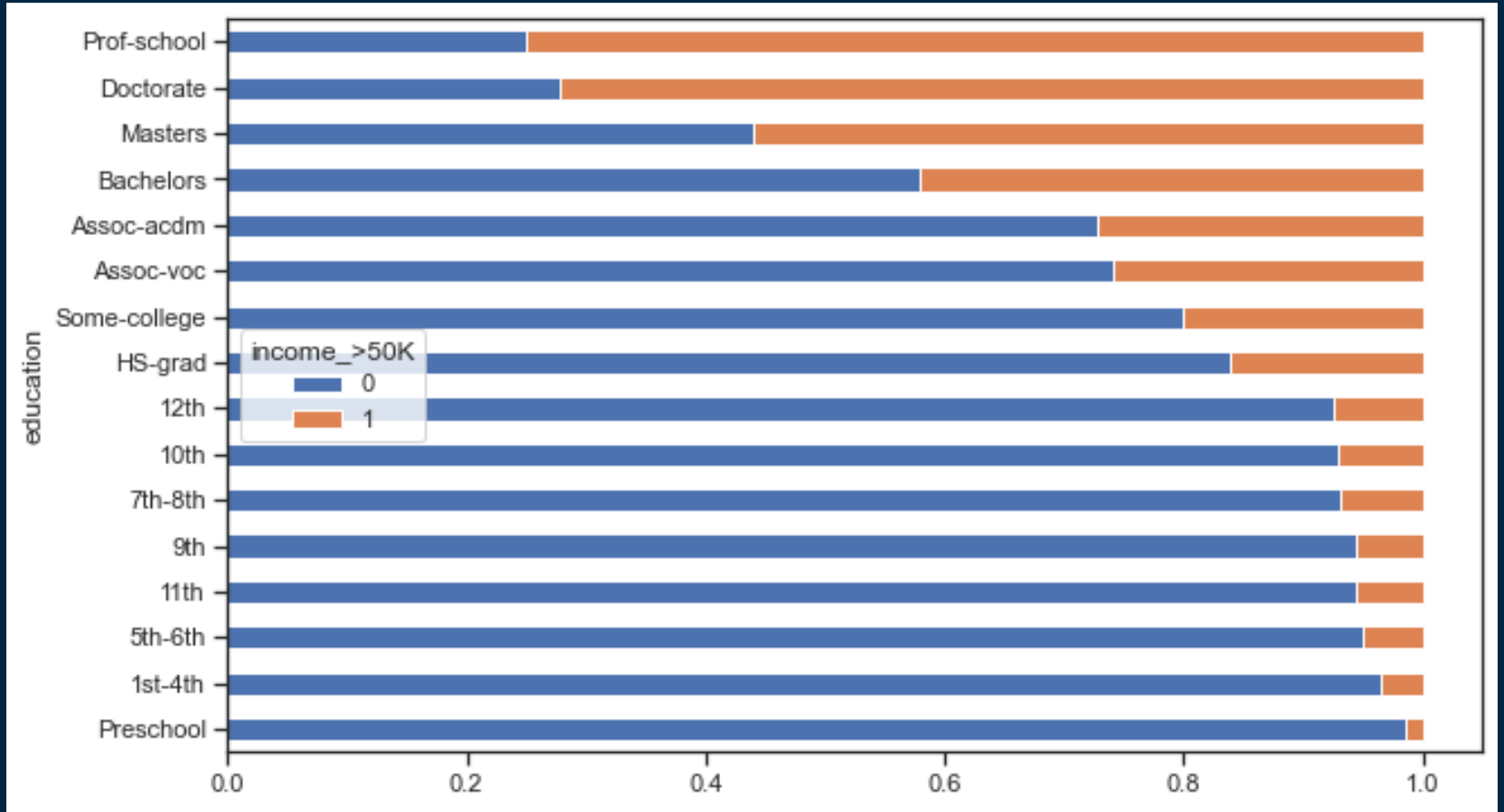
بررسی داده‌های ناشناخته

- حدود ۳۰۰۰ رکورد دارای مقدار nan
- ۲۵۰۰ رکورد شامل دو مقدار nan (workclass - occupation)
- درصد پایینی از داده ناشناخته است
- نتیجه نهایی – حذف داده های ناشناخته

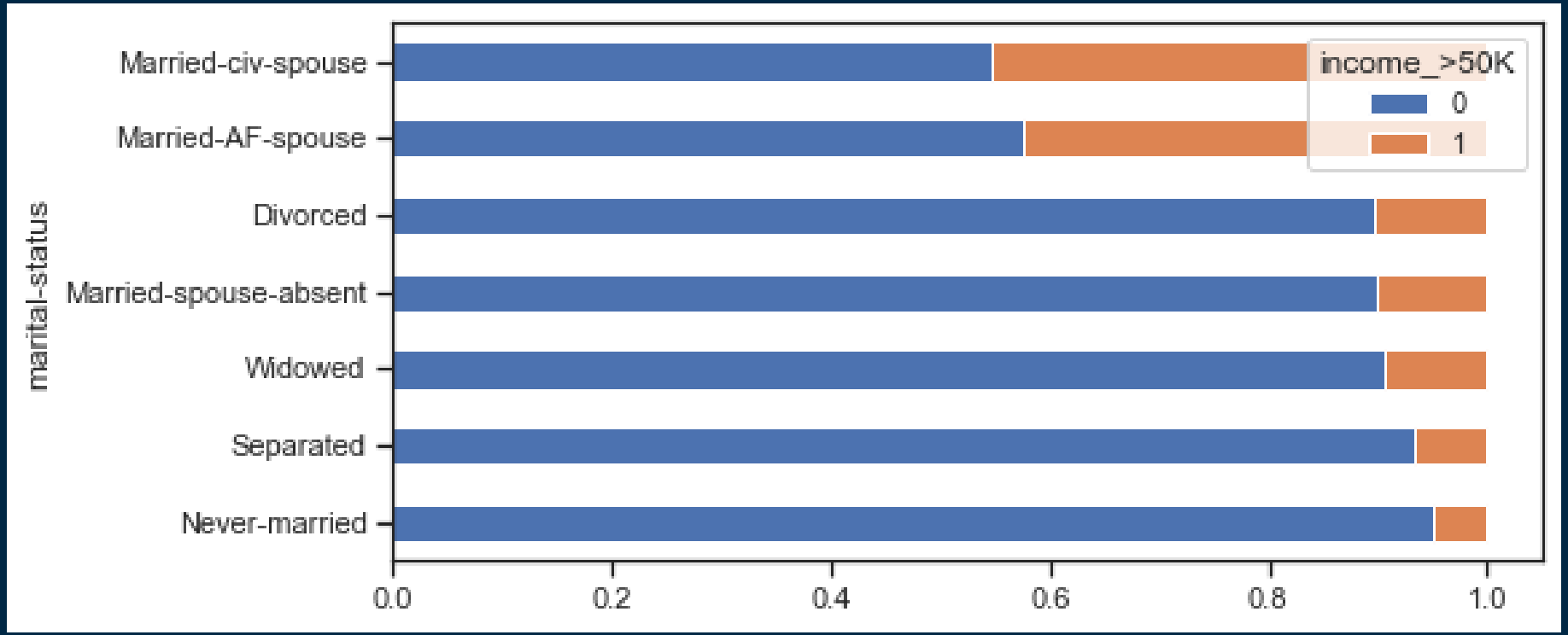
بررسی جزئی خصیصه‌های داده



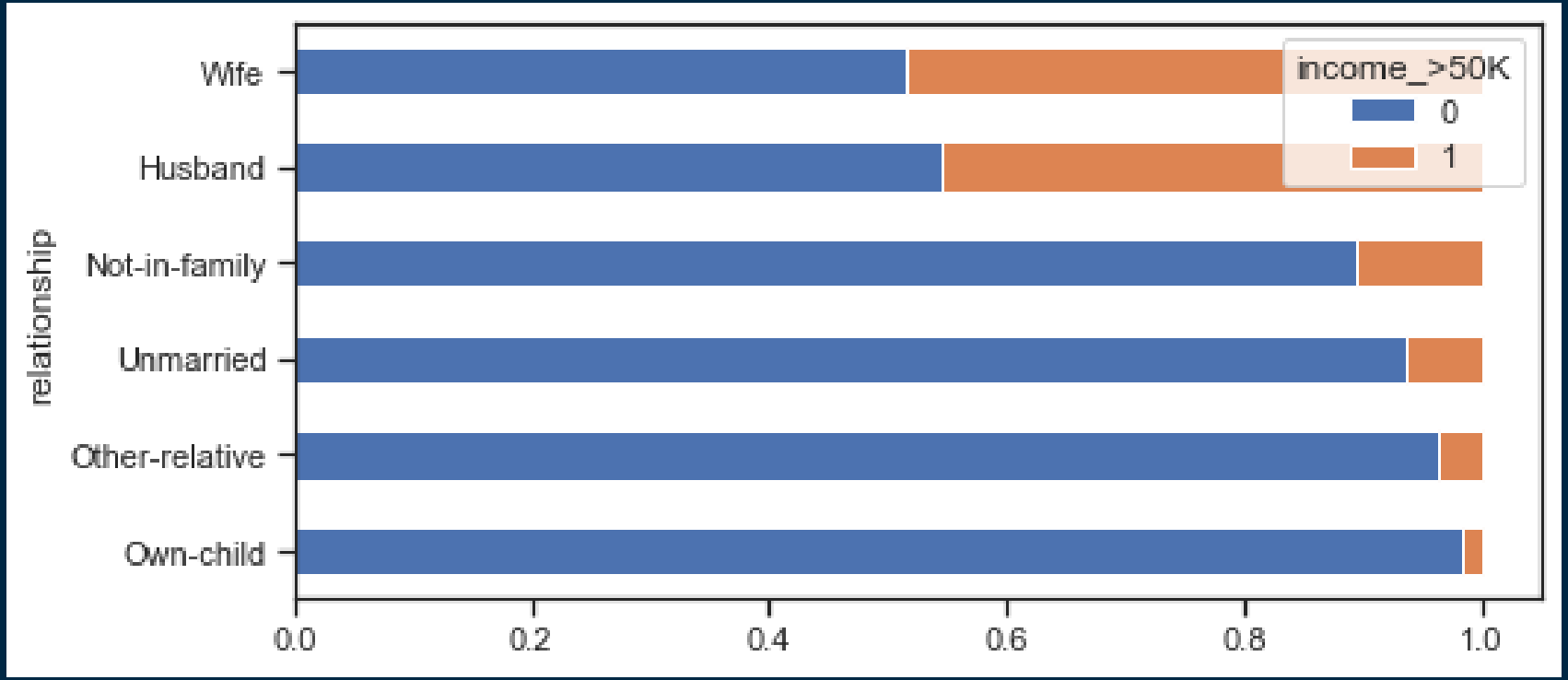
بررسی جزئی خصیصه‌های داده



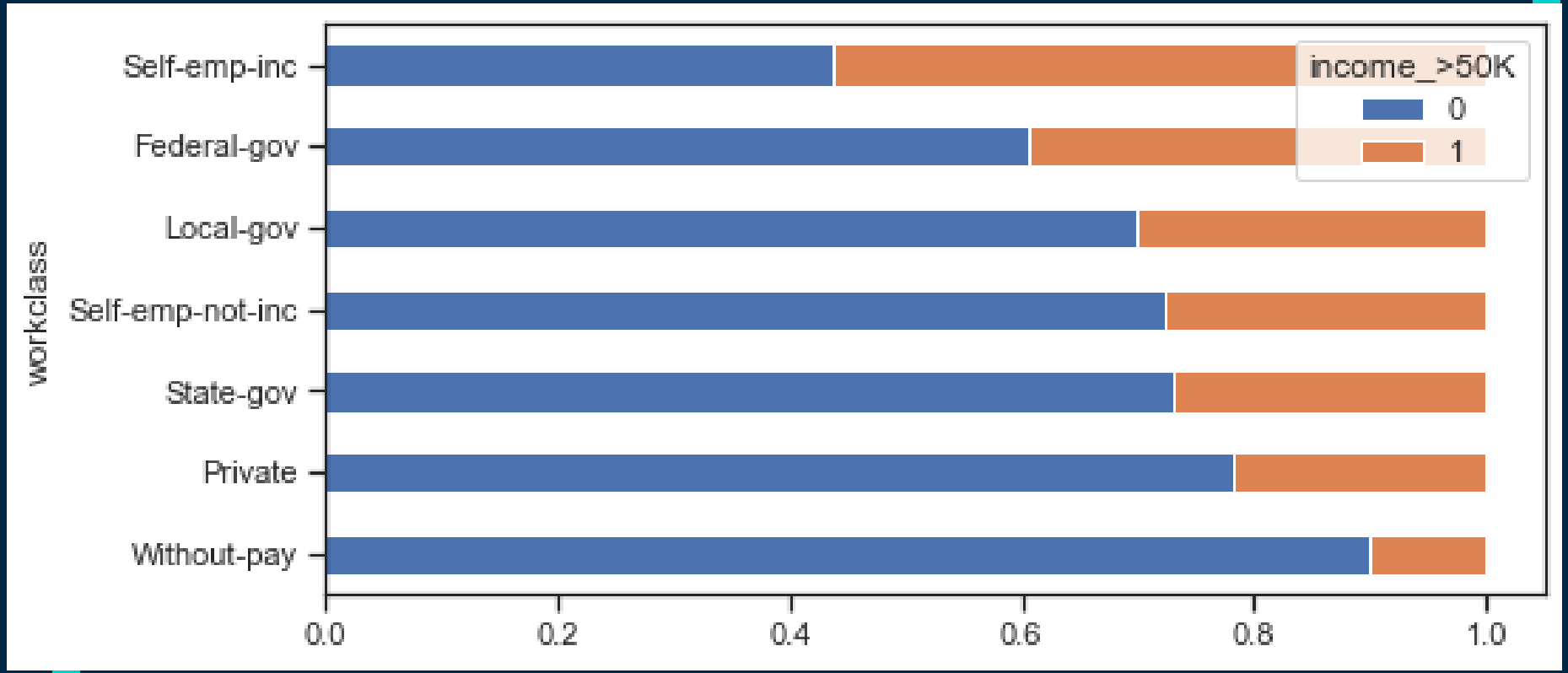
بررسی جزئی خصیصه‌های داده



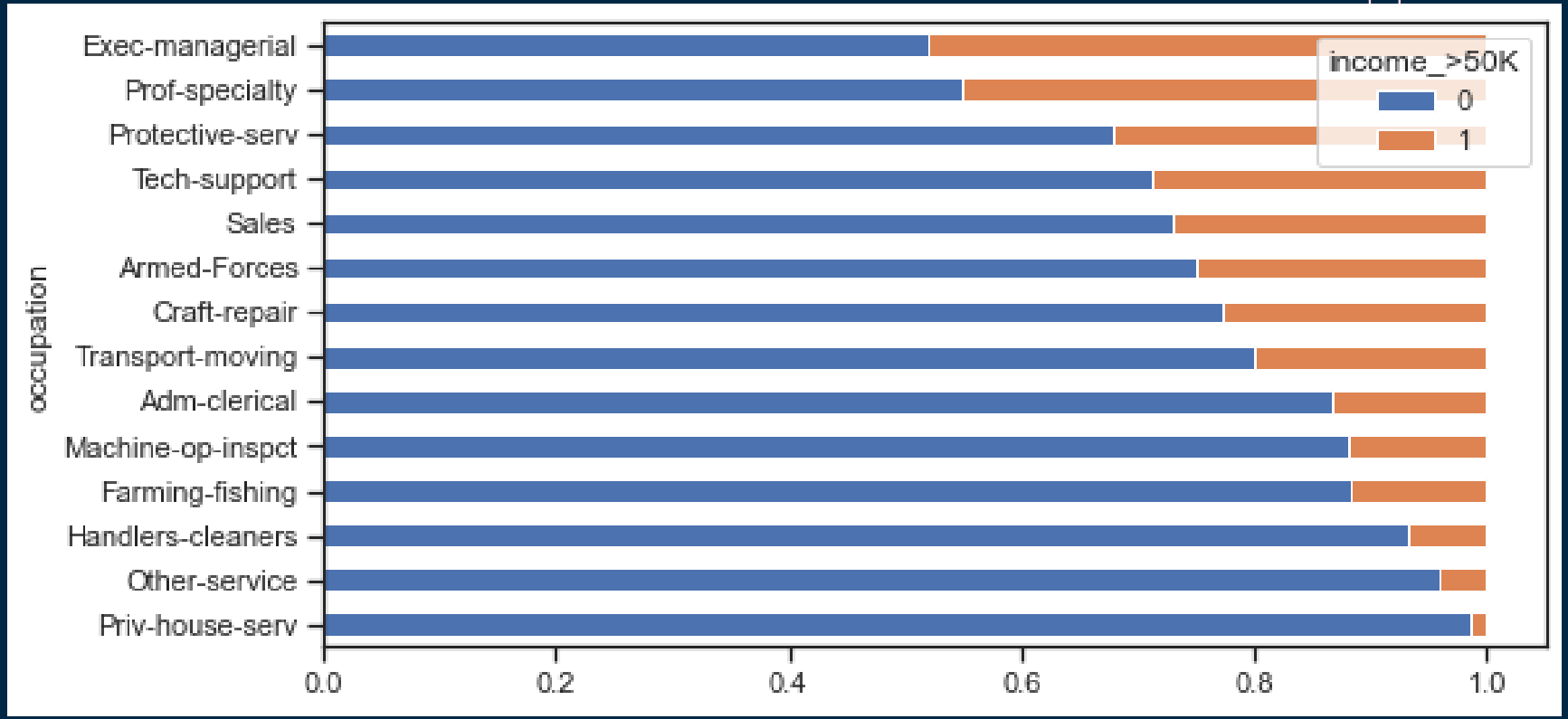
بررسی جزئی خصیصه‌های داده



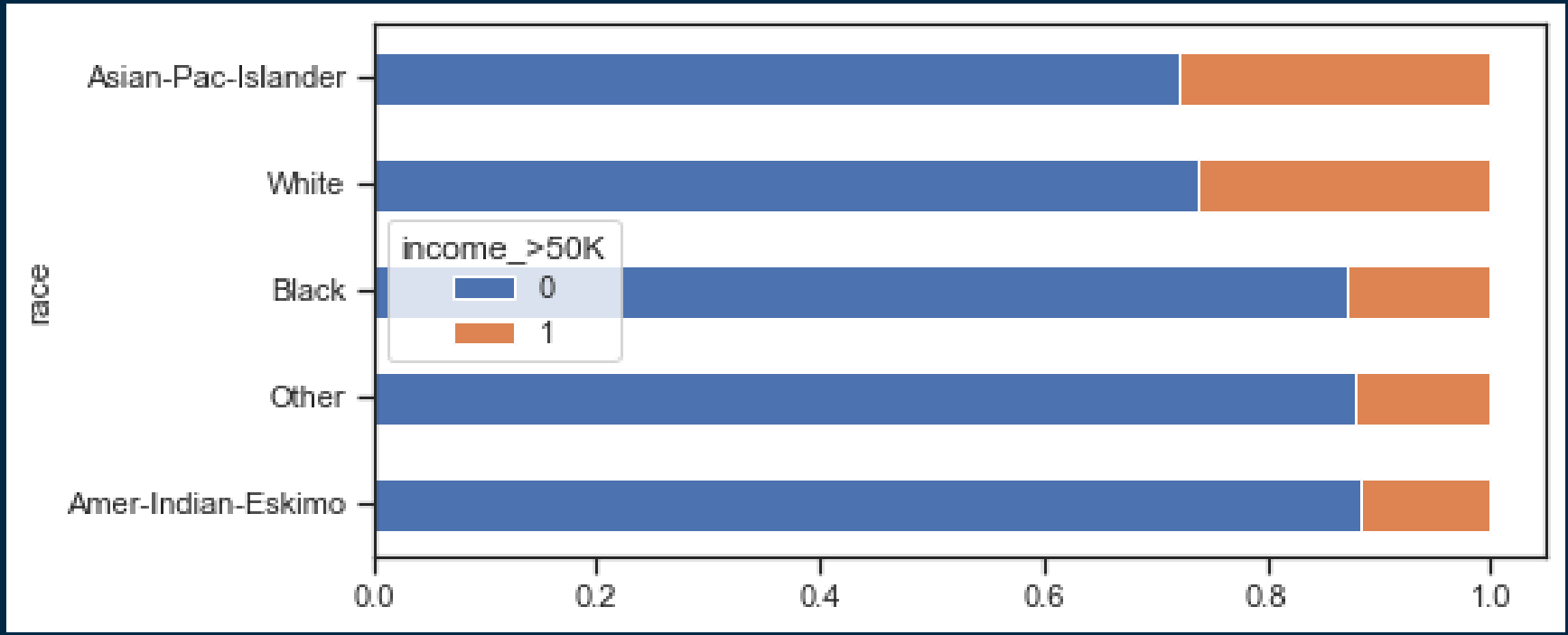
بررسی جزئی خصیصه‌های داده



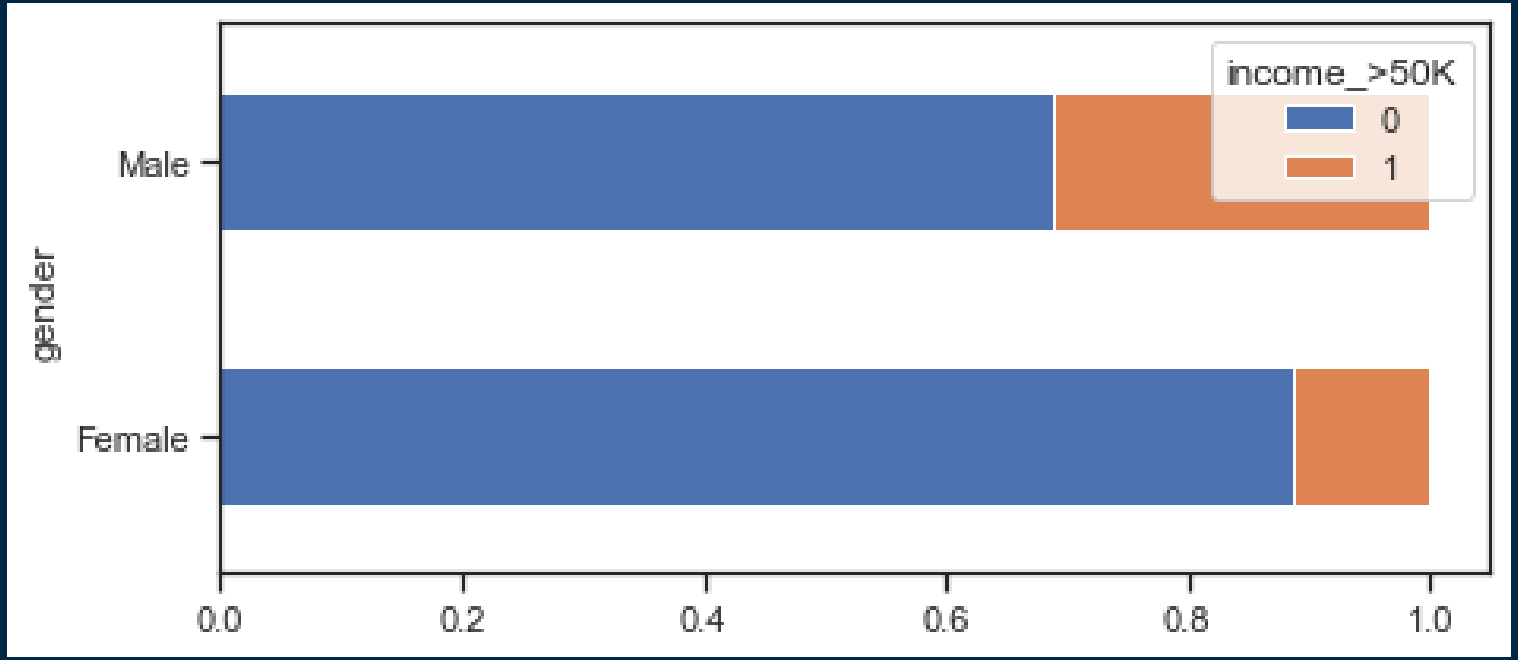
بررسی جزئی خصیصه‌های داده



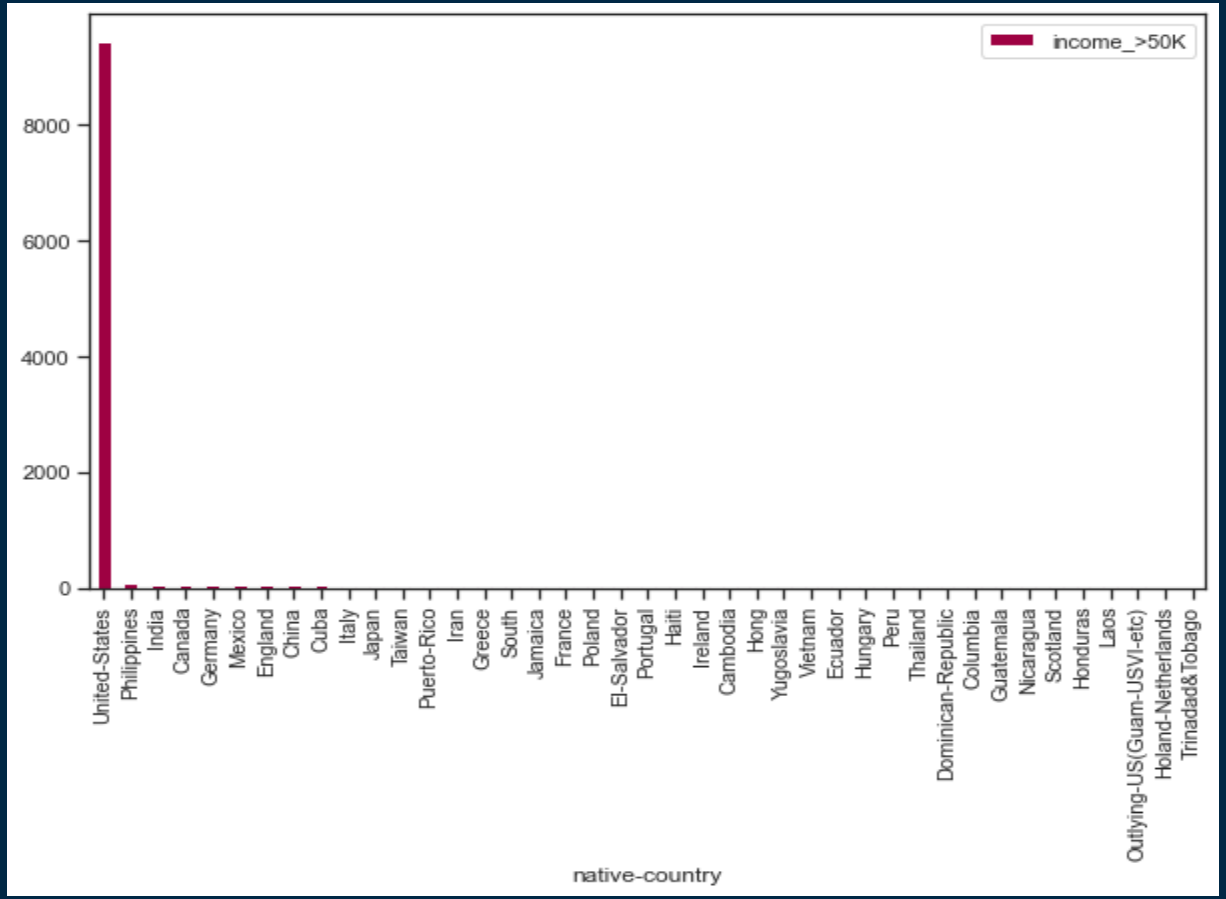
بررسی جزئی خصیصه‌های داده



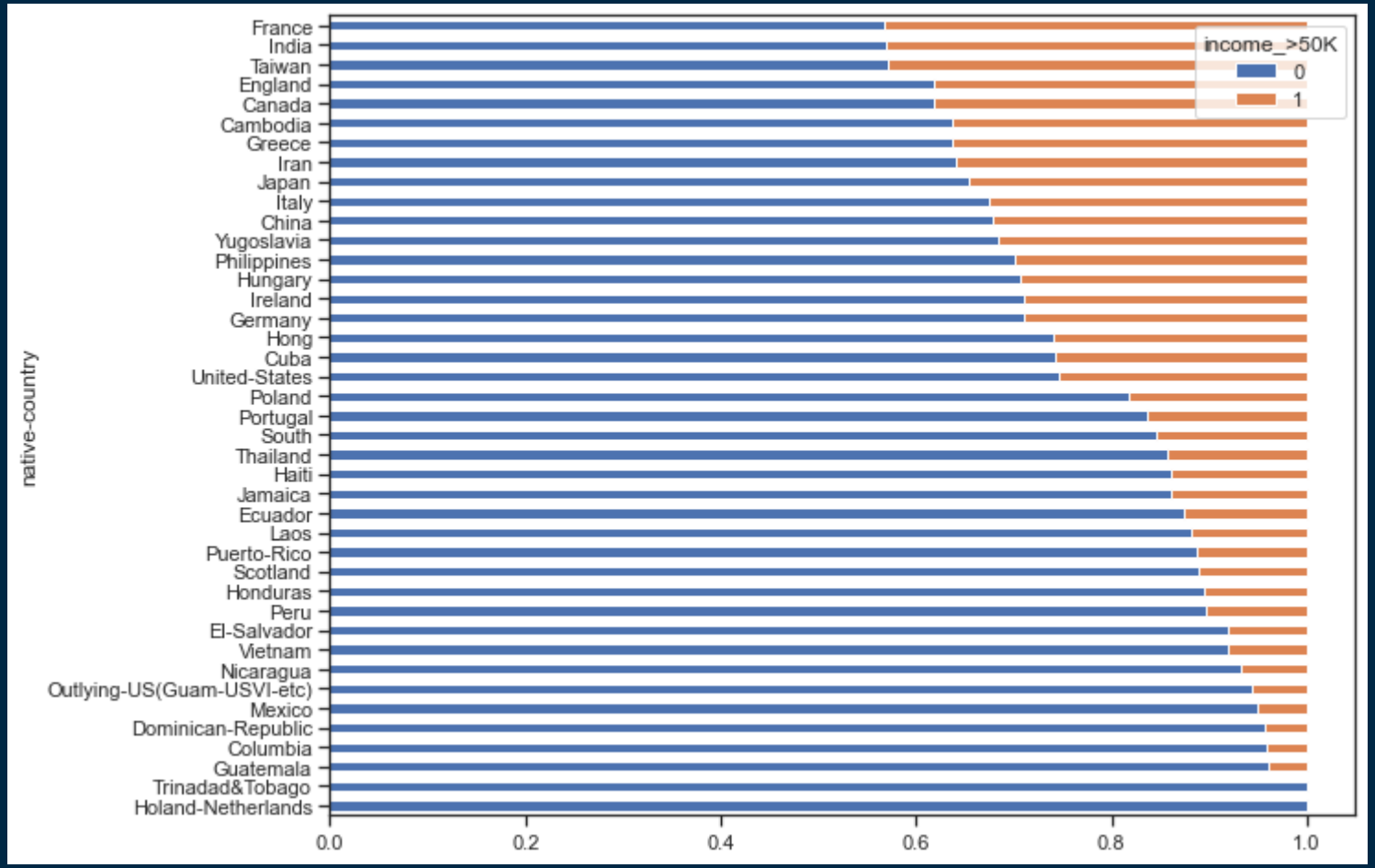
بررسی جزئی خصیصه‌های داده



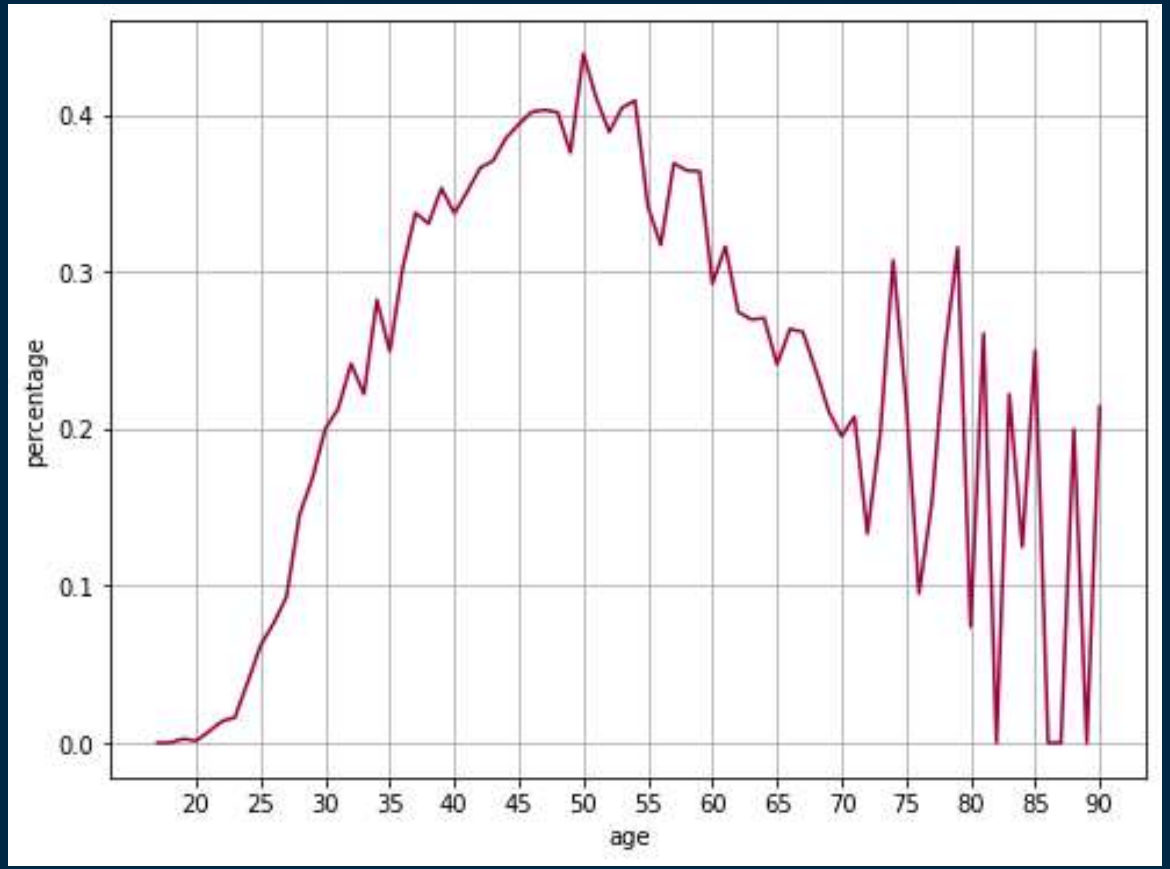
بررسی جزئی خصیصه‌های داده



بررسی جزئی خصیصه‌های داده

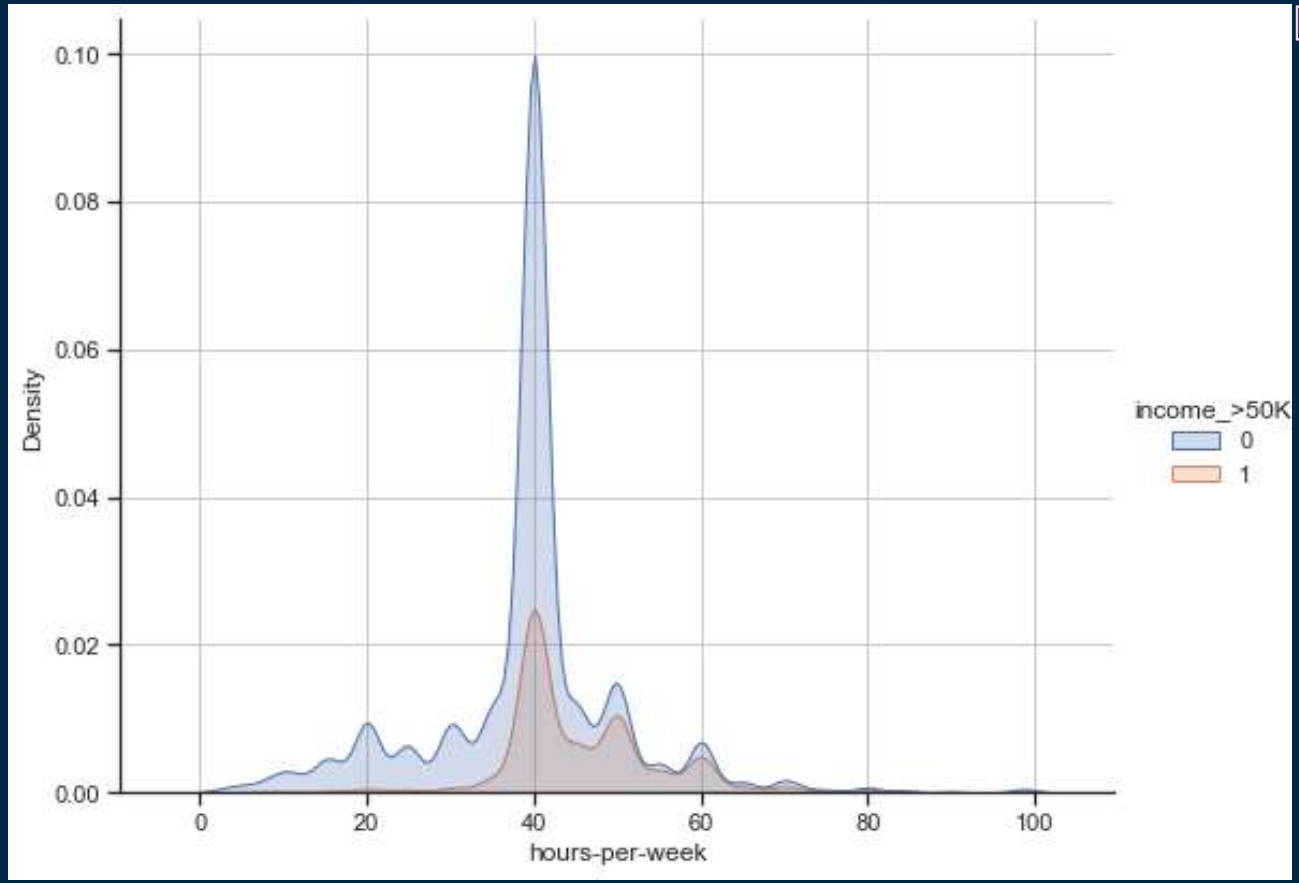


بررسی جزئی خصیصه‌های داده

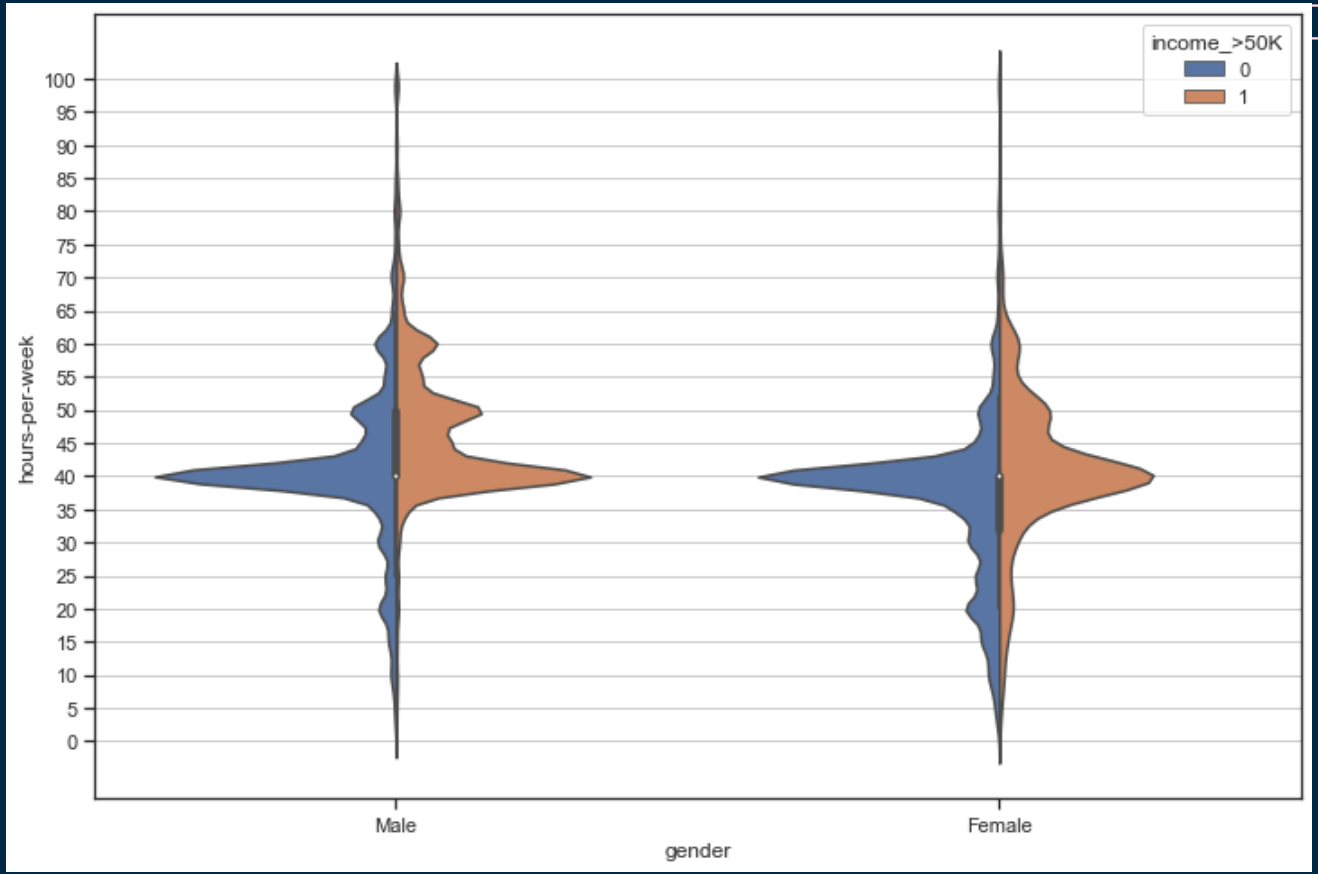


توزیع افراد پردرآمد
نسبت به سن

بررسی جزئی خصیصه‌های داده



بررسی جزئی خصیصه‌های داده



فهرست مطالب

آزمایشات

حل مسئله

انکد ویژگی‌های categorical [3]

- استفاده از روش MCA [4]
  Multiple Correspondence Analysis
- جلوگیری از نفرین ابعاد
- به دست آوردن بهترین ترکیب ممکن از ویژگی‌های categorical
- مشابه روش PCA
- انکد one-hot بر ستون gender

انتخاب ویژگی

- استفاده از دو معیار
  - ANOVA F-value
  - mutual information
- انتخاب ۳ ویژگی ضعیف
- حذف ویژگی مشترک اعلام شده از این دو معیار – fnlwgt

تبدیل داده (transform)

- ذخیره داده‌های تبدیل نشده برای استفاده در random forest
- نرمال کردن همه ویژگی‌ها به جز MCA_x و gender_Male
- ذخیره ۱۵ درصد داده به عنوان داده تست
- عدم گسسته کردن داده

انتخاب و آموزش مدل

- ۷۵ درصد دقت به عنوان baseline
- مدل random forest
- مدل K-NN

انتخاب و آموزش مدل (random forest)

- استفاده از Randomized Search CV
  - n_estimators
  - max_depth
  - min_samples_split
  - min_samples_leaf
  - criterion
- 100 iterations
- 5-fold

انتخاب و آموزش مدل (K-NN)

- استفاده از Grid Search CV
  - n_neighbors

- 5-fold

- استفاده از داده‌های transform شده در این مدل

فهرست مطالب

نتایج

نتیجه تست

- استفاده از داده جدا شده برای تست

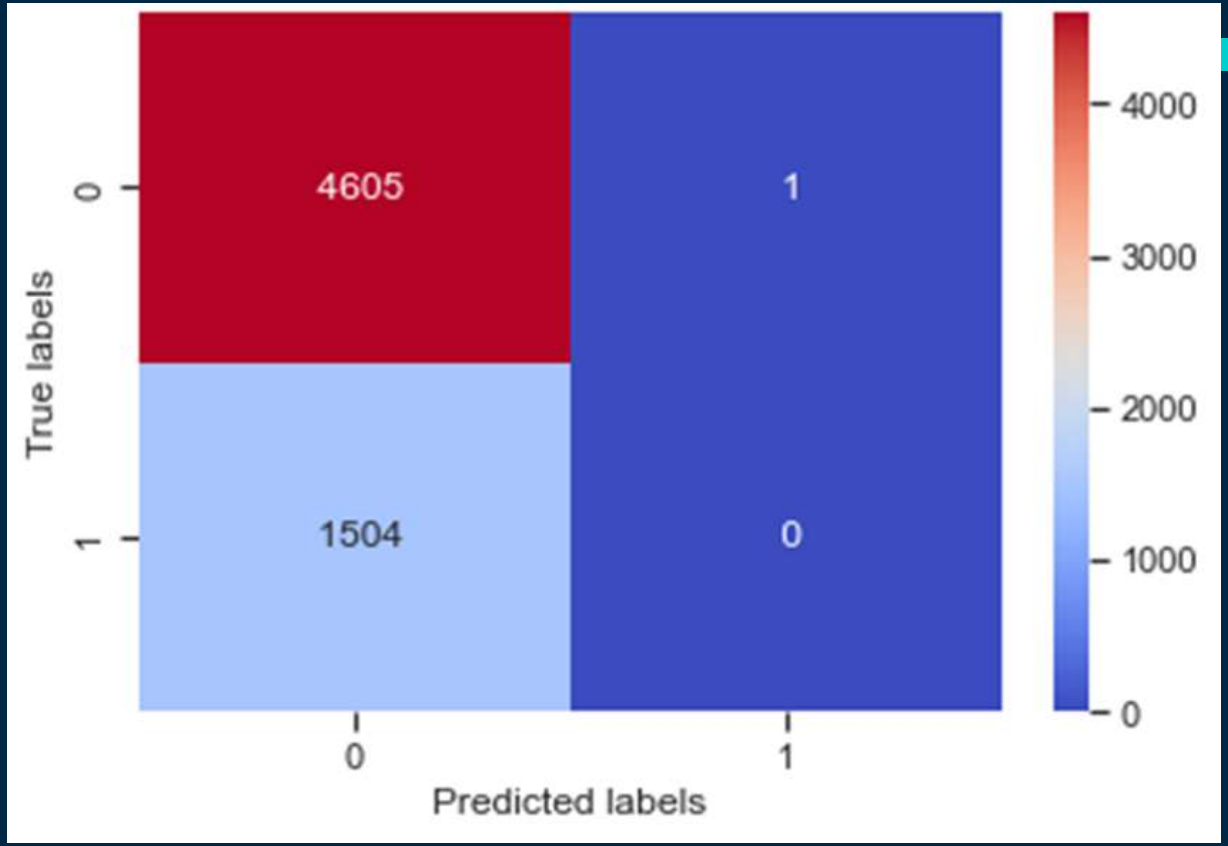
- مقایسه دو مدل بر اساس معیارهای
  - Baseline accuracy
  - Precision
  - Recall
  - F-score

نتیجه تست (random forest)

Accuracy score: 75%

Precision class 0: 75%
Recall class 0: 99%
F-score class 0: 85%

Precision class 1: 0%
Recall class 1: 0%
F-score class 1: 0%

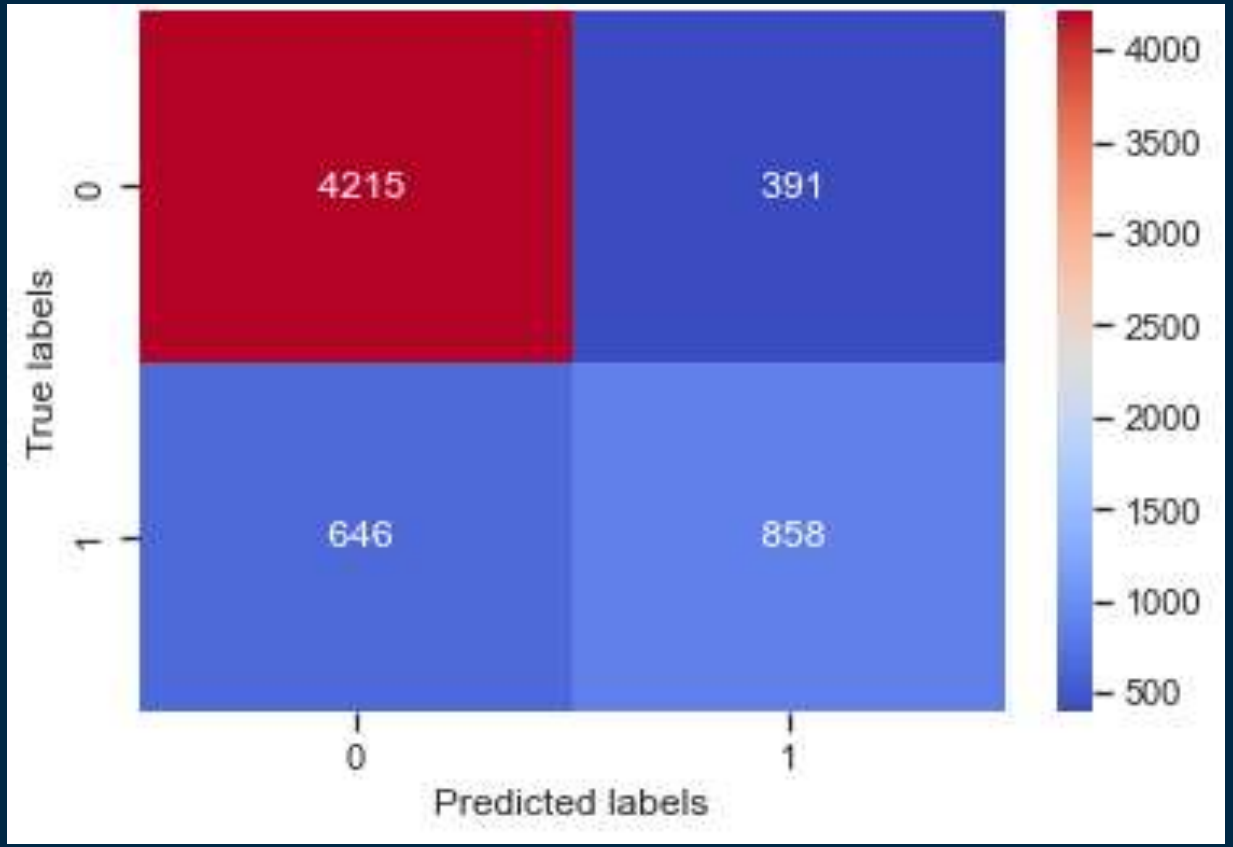


نتیجه تست (K-NN)

Accuracy score: 83%

Precision class 0: 86%
Recall class 0: 91%
F-score class 0: 89%

Precision class 1: 68%
Recall class 1: 57%
F-score class 1: 62%



انتخاب مدل برتر

- عملکرد ضعیف random forest
- تشخیص کلاس 1 توسط K-NN
- مدل برتر – K-NN

فهرست مطالب

پیشنهادات

پیشنهادات

- استفاده از مدل‌های دیگر در آزمایشات
- استفاده از انکدهای دیگر برای داده‌های categorical برای مدل random forest
- بررسی عملکرد روش‌های مختلف Transform داده

فهرست مطالب

نتیجه گیری

نتیجه‌گیری

- بررسی نوع مسئله و داده در دست در ابتدای پروژه
  - مسئله binary classification
  - داده سرشماری وضعیت درآمد
- بررسی ویژگی های داده
  - حذف داده‌های ناشناخته
  - بررسی جزئی ویژگی‌های داده

نتیجه‌گیری

- بررسی مسئله و حل آن
  - استفاده از MCA برای انکد داده های categorical
  - انتخاب و حذف ویژگی ضعیف با استفاده از دو معیار mutual information و ANOVA F-value
  - تبدیل داده با استفاده از روش نرمالایز و ذخیره ۱۵ درصد داده برای تست
  - آموزش دو مدل random forest و K-NN
- انتخاب مدل
  - انتخاب مدل K-NN بر اساس معیارهای baseline accuracy و F-Score

فهرست مطالب

مراجع

1) https://archive.ics.uci.edu/ml/datasets/adult

2) https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names

3) https://maxhalford.github.io/prince/mca/

4) https://www.ibm.com/docs/he/spss-statistics/25.0.0?topic=categories-multiple-correspondence-analysis