Agenda

- Objective
- Data exploration
- Data preprocessing
- Model evaluation
- Model hyperparameter findings

Objective

- Predict person's income > $50,000 / year

- Target is to find income
  - > $50,000 / year -> 1
  - <= $50,000 / year -> 0

Data exploration

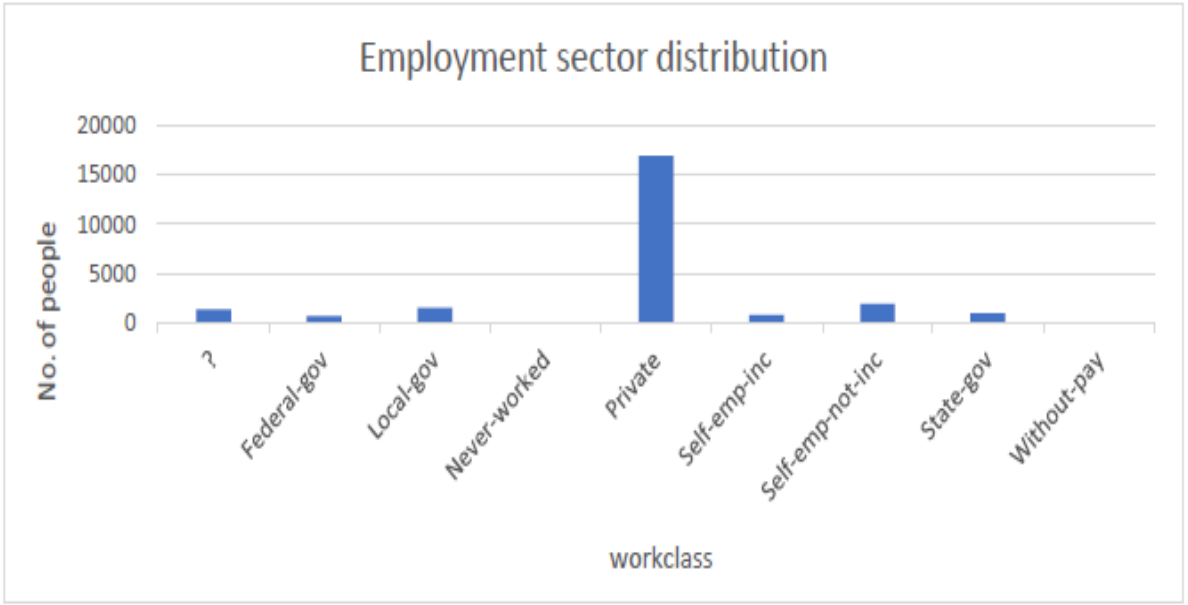
- 13 features (exclude target)

- 6 numerical
- 7 categorical

Features	Data type	Example value
age	Numerical	60
workclass	Categorical	Private
fnlwgt	Numerical	173960
education	Categorical	Bachelors
education-num	Numerical	13
marital-status	Categorical	Divorced
occupation	Categorical	Prof-specialty
relationship	Categorical	Not-in-family
sex	Categorical	Female
capital-gain	Numerical	0
capital-loss	Numerical	0
hours-per-week	Numerical	42
native-country	Categorical	United-States

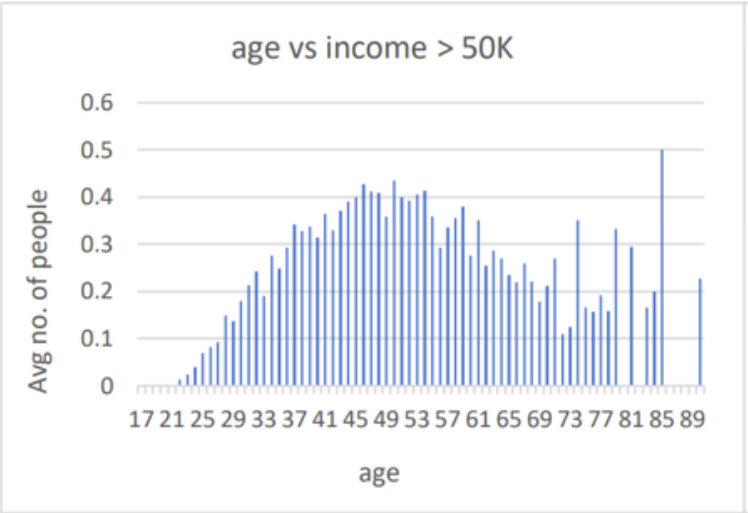
Data exploration - workclass

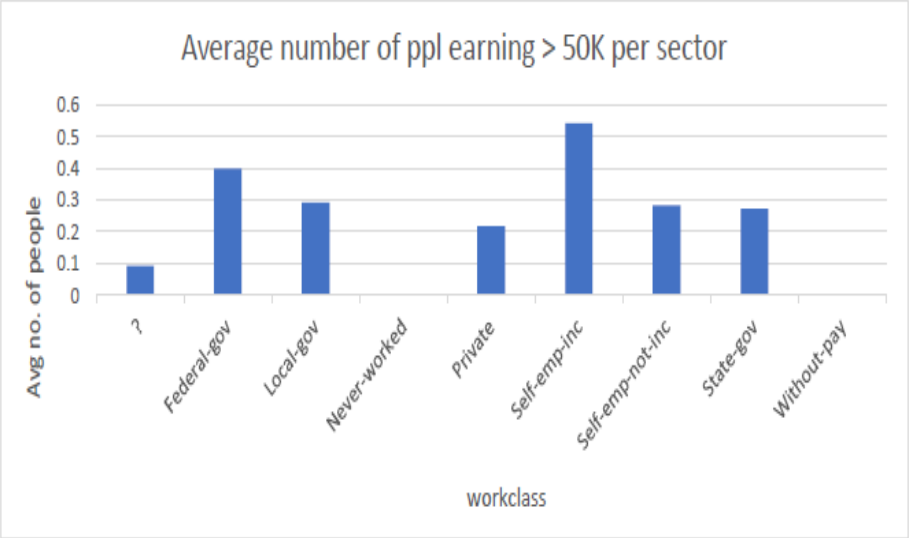
- Workclass = employment sector
- Each label have diff. number of records belongs to
- Use average no. of ppl earning > 50k as metric



Data exploration - reasonable distribution

- Take avg. no. of ppl earning > 50k
- 7 features has reasonable distribution on value vs earnings
  - Workclass
  - Age
  - Marital-status
  - Occupation
  - Sex
  - Hours-per-week
  - Native-country
- Use as training features



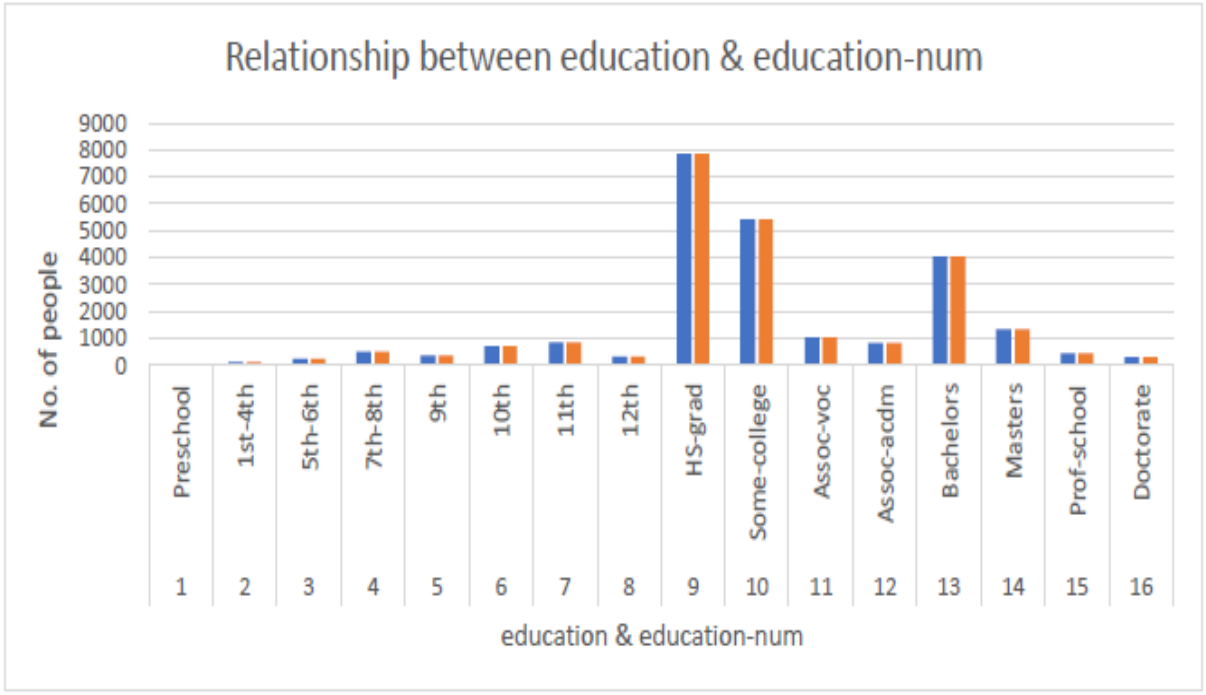


Data exploration - education vs education_num

- 2 separate features
- Education = in wordings
- Education_num = in values

- Natural ranking ordering
- Label encoding

- Keep education_num only

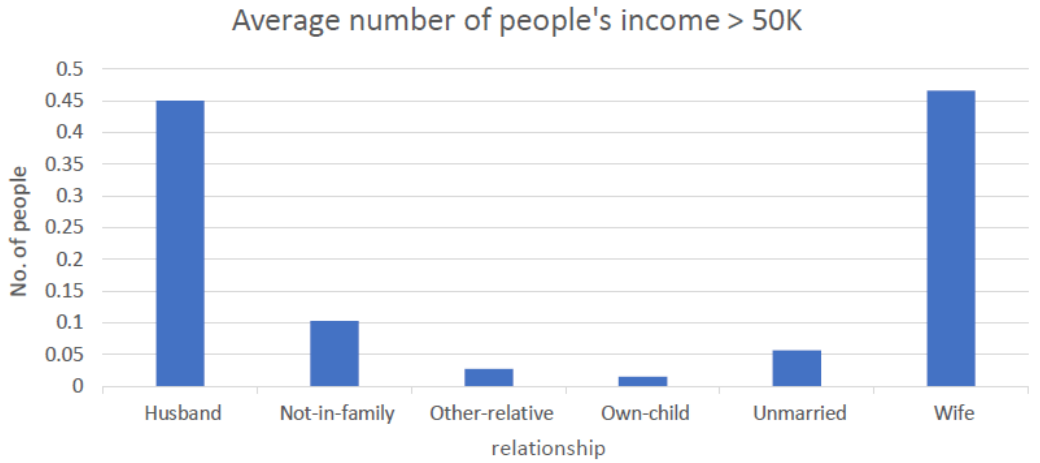


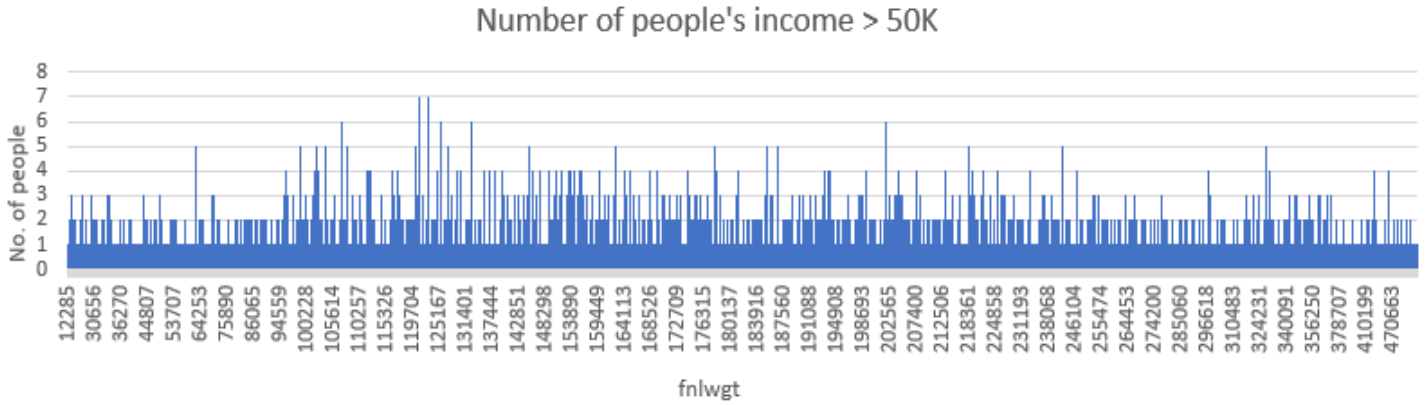
Data exploration - test & drop features

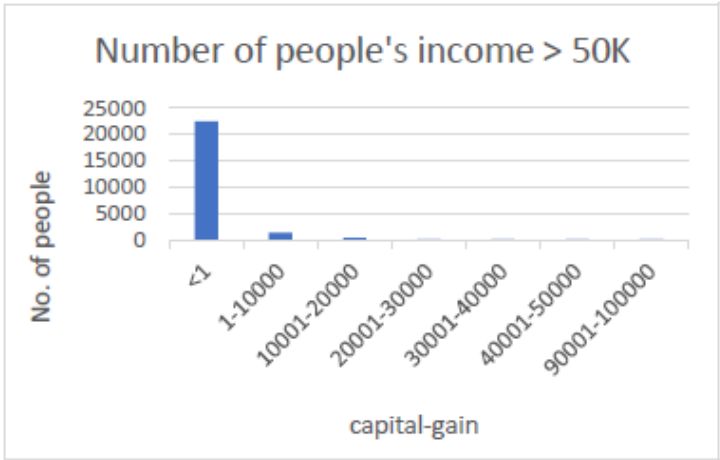
- Quite stable across label
- Extreme distribution

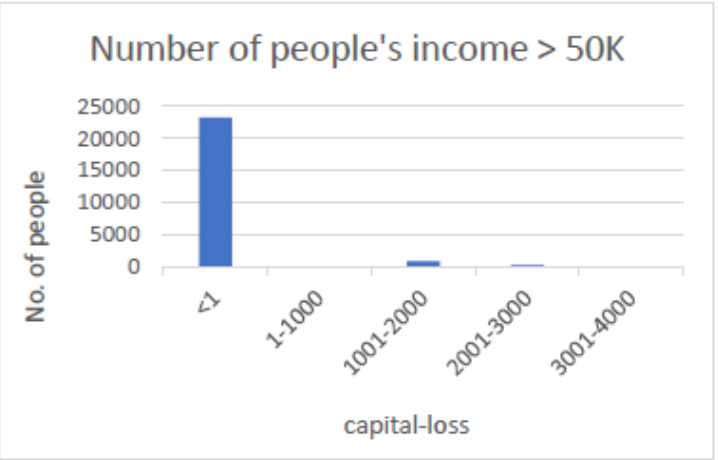
- Test is it help for prediction
  - In terms of F1 score

- Drop
  - fnlwgt
  - relationship
- Keep
  - capital-gain
  - capital-loss

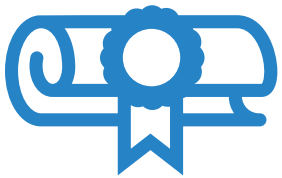








Data preprocessing - splitting data



Training dataset (train.csv)

Split to 70% for training, 30 % for testing

Denoted as training & testing data

Testing dataset (test.csv)

Predict and submit to Kaggle

Denoted as evaluation data

Data preprocessing - cleaning

- Remove
  - education
  - fnlwgt
  - relationship
- Remove extra white spaces of value
  - E.g. “Male ”

Data preprocessing - label encoding

- Convert the label into a numeric form
  - by natural ranking ordering
- Features transformed
  - Education -> Education-num
  - Sex

Label (education)	Numeric form (education-num)
11th	7
12th	8
HS-grad	9
Some-college	10

Label (sex)	Numeric form (sex)
Male	0
Female	1

Data preprocessing - one hot encoding

- Convert categorical data into multiple columns depending on the categories in a column.
- Only contain 1 or 0 corresponding to the categorical label
- Features transformed
  - Workclass
  - Marital-status
  - Occupation
  - Native-country

Before encoding
Private
Self-emp-inc
State-gov

Encode ->

Private	Self-emp-inc	State-gov
1	0	0
0	1	0
0	0	1

Data preprocessing - encoding & normalization

- Transform both training and testing dataset
- Only training dataset fit & transform encoding model
- Testing & Evaluation dataset transform only

- Lastly, normalize all features value

Model selection & evaluation

- Evaluation metric = F1 score
  - Higher better
- In total 14 models tried

#	Model	F1 score	Accuracy
10	XGBClassifier(eta=0.2)	0.7052	0.8732
12	XGBClassifier(eta=0.3)	0.7048	0.8718
11	XGBClassifier(eta=0.1)	0.7032	0.8751
7	RandomForestClassifier(random_state=30, n_estimators=120, min_samples_leaf=2, class_weight=balanced)	0.7022	0.8376
13	XGBClassifier(eta=0.2, learning_rate=0.05, min_child_weight=3, max_depth=10)	0.6976	0.8735
9	GradientBoostingClassifier	0.676	0.8673

#	Model	F1 score	Accuracy
8	AdaBoostClassifier	0.6658	0.862
4	make_pipeline(StandardScaler(), SVC(kernel=linear))	0.6537	0.8545
5	LogisticRegression(penalty=none)	0.6533	0.8534
6	RandomForestClassifier	0.6467	0.8441
14	Neural Network	0.6377	0.8351
3	make_pipeline(StandardScaler(), SVC(gamma=auto))	0.6352	0.8501
2	DecisionTreeClassifier	0.6109	0.8197
1	GaussianNB	0.4451	0.4367

Model hyperparameter findings

Model with custom parameter (#7)

>

Model without custom parameter (#6)

#	Model	F1 score	Accuracy
7	RandomForestClassifier(random_state=30, n_estimators=120, min_samples_leaf=2, class_weight=balanced)	0.7022	0.8376
6	RandomForestClassifier	0.6467	0.8441

#	Model	F1 score	Accuracy
4	make_pipeline(StandardScaler(), SVC(kernel=linear))	0.6537	0.8545
3	make_pipeline(StandardScaler(), SVC(gamma=auto))	0.6352	0.8501

Model with simpler parameter setting (#10, #11, #12)

>

Model with complex parameter setting (#13)

#	Model	F1 score	Accuracy
10	XGBClassifier(eta=0.2)	0.7052	0.8732
12	XGBClassifier(eta=0.3)	0.7048	0.8718
11	XGBClassifier(eta=0.1)	0.7032	0.8751
13	XGBClassifier(eta=0.2, learning_rate=0.05, min_child_weight=3, max_depth=10)	0.6976	0.8735