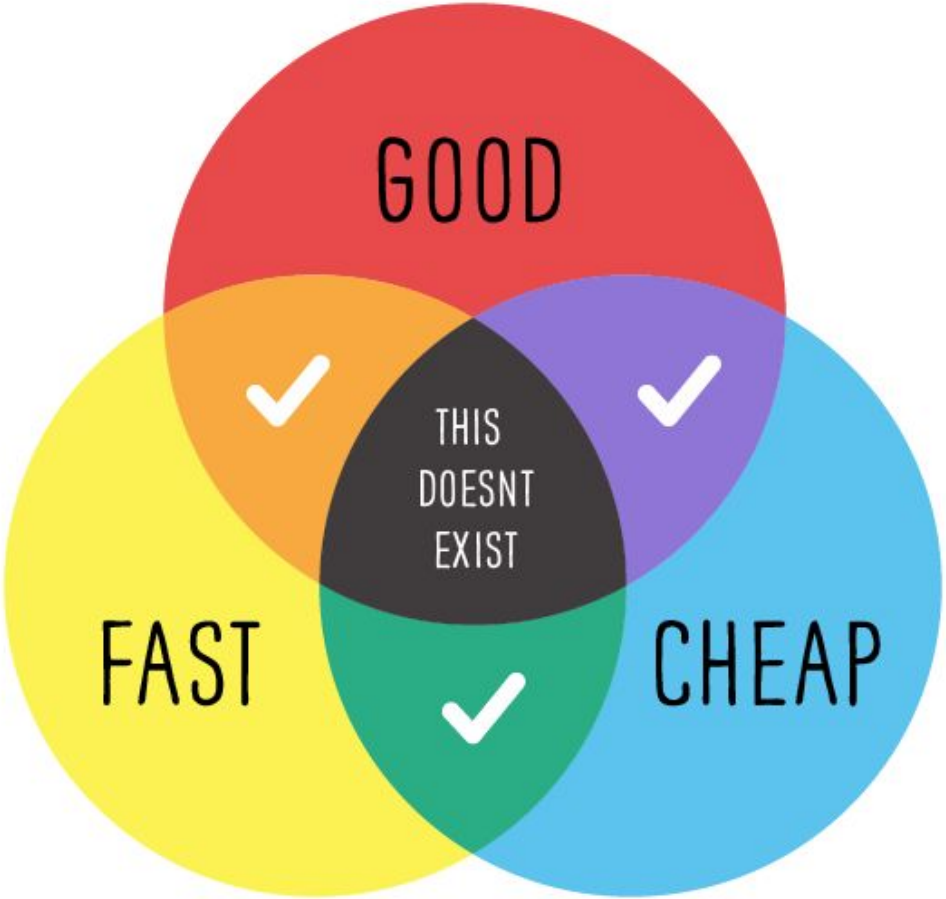
Wage Prediction

Good-Fast-Cheap Challenge

Billy & Bird Sr.



Our Team Constraint

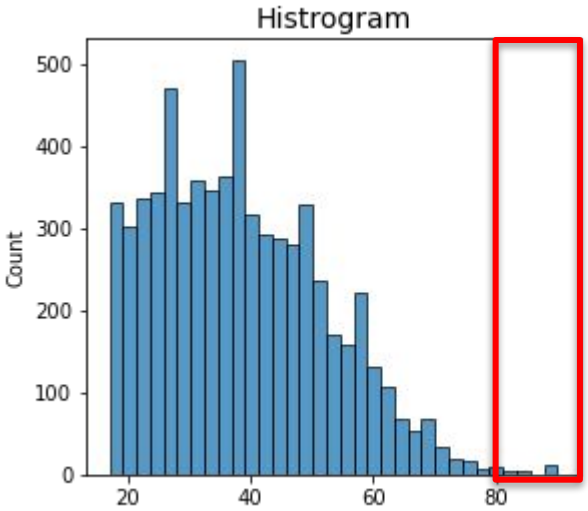
- Your choice of algorithm
- Your choice of features
- **Must use the cheap train sample**

Our data : 6,513 rows × 13 features

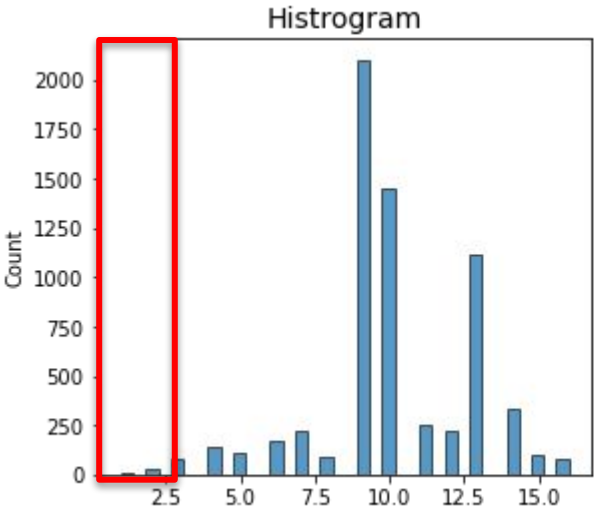
HEINLEY

Understand our data

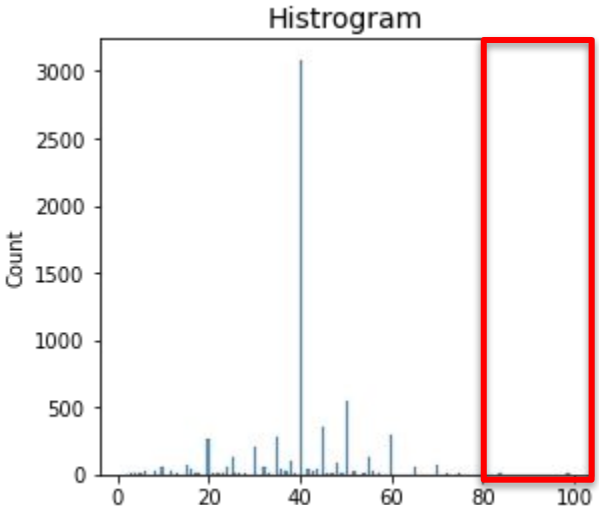
EDA: Outliers



Age

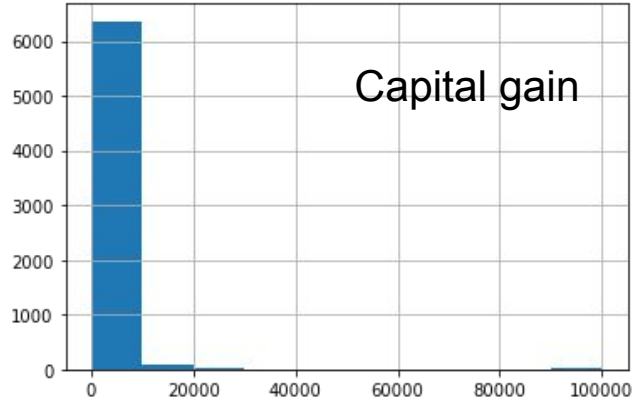


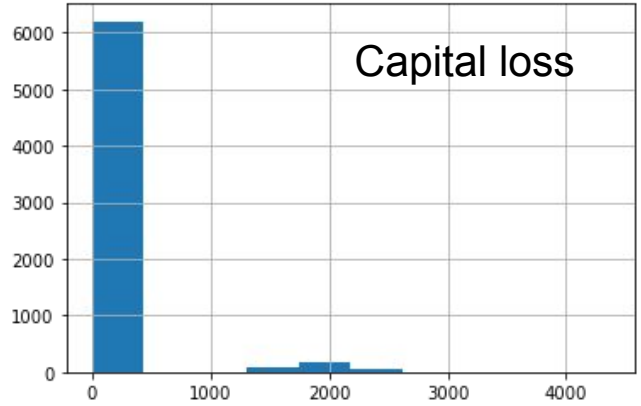
Education-num

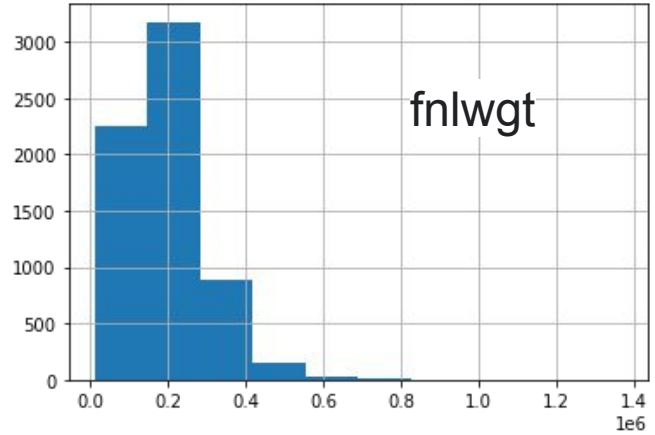


Hour-per-week

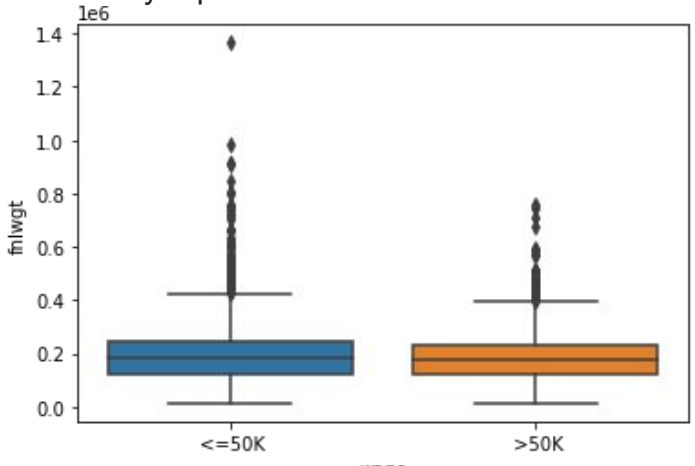
Feature Selection: Excluded







final weight. In other words, this is the number of people the census believes the entry represents



Feature Selection: duplicated features

Education	
HS-grad	2103
Some-college	1451
Bachelors	1113
Masters	334
Assoc-voc	250
11th	225
Assoc-acdm	222
10th	175
7th-8th	142
9th	106
Prof-school	103
12th	89
Doctorate	81
5th-6th	79
1st-4th	27
Preschool	13

=

Education-num	
9	2103
10	1451
13	1113
14	334
11	250
7	225
12	222
6	175
4	142
5	106
15	103
8	89
16	81
3	79
2	27
1	13

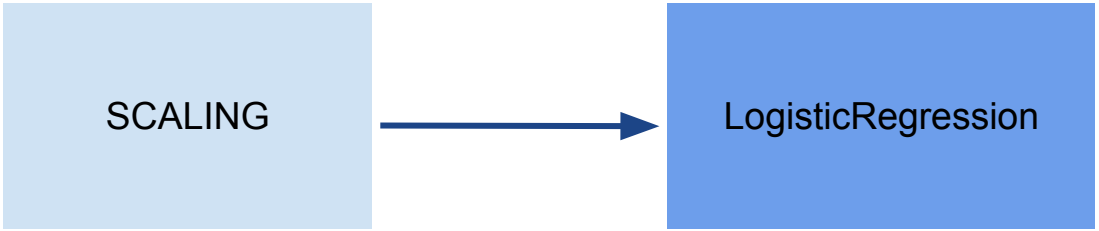
Ordinal Var. ⇒ Dummy

Feature Engineering

United-States	5807
Mexico	150
?	120
Philippines	43
El-Salvador	29
Canada	26
Germany	24
Dominican-Republic	20
South	18
Jamaica	18
India	17
China	17
Puerto-Rico	17
England	16
Cuba	16
Italy	15
Poland	14
Japan	12
Haiti	11

Too many categories, group into **Other**

First Trial-run Model



Dataset	Accuracy
Train	0.8492
Test	0.8494

Classes	F1 - score
<= 50 K	0.90
> 50 K	0.64

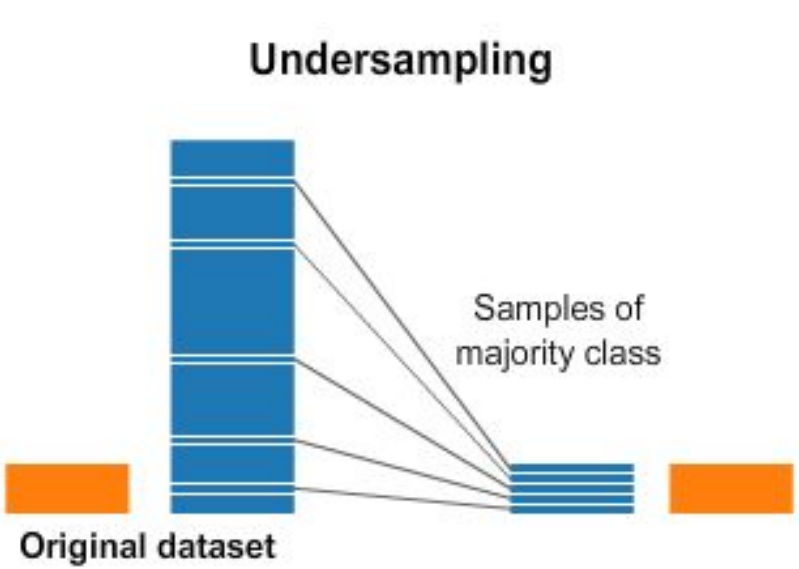
Modeling

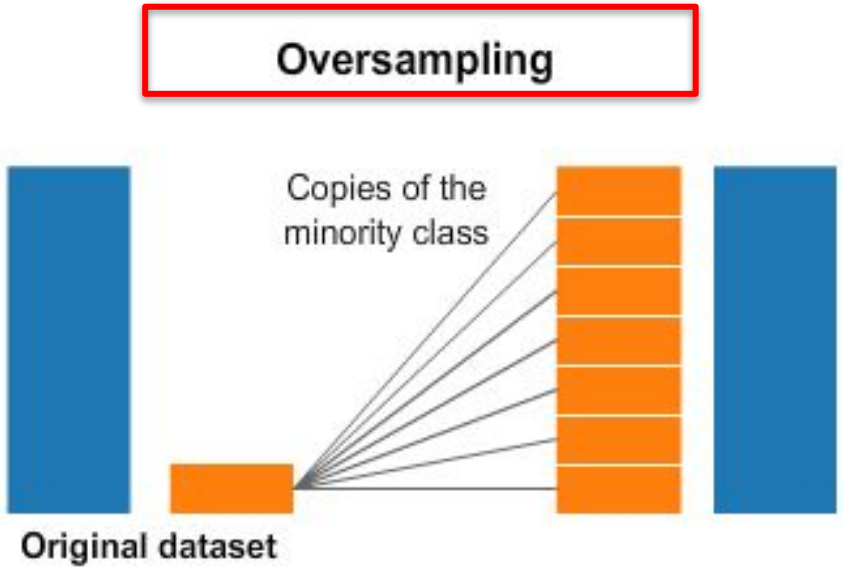
- Handling the regularization
- Handling the imbalance classes
- Fast & Cheap algo.
- Regularization - ‘L2’

Baseline score:

0	75.98
1	24.02

Dealing with Imbalanced Classes





SMOTE : Wage <=50K 75.98 % ===> 50%
Wage >50K 24.02 % ===> 50%

Evaluation

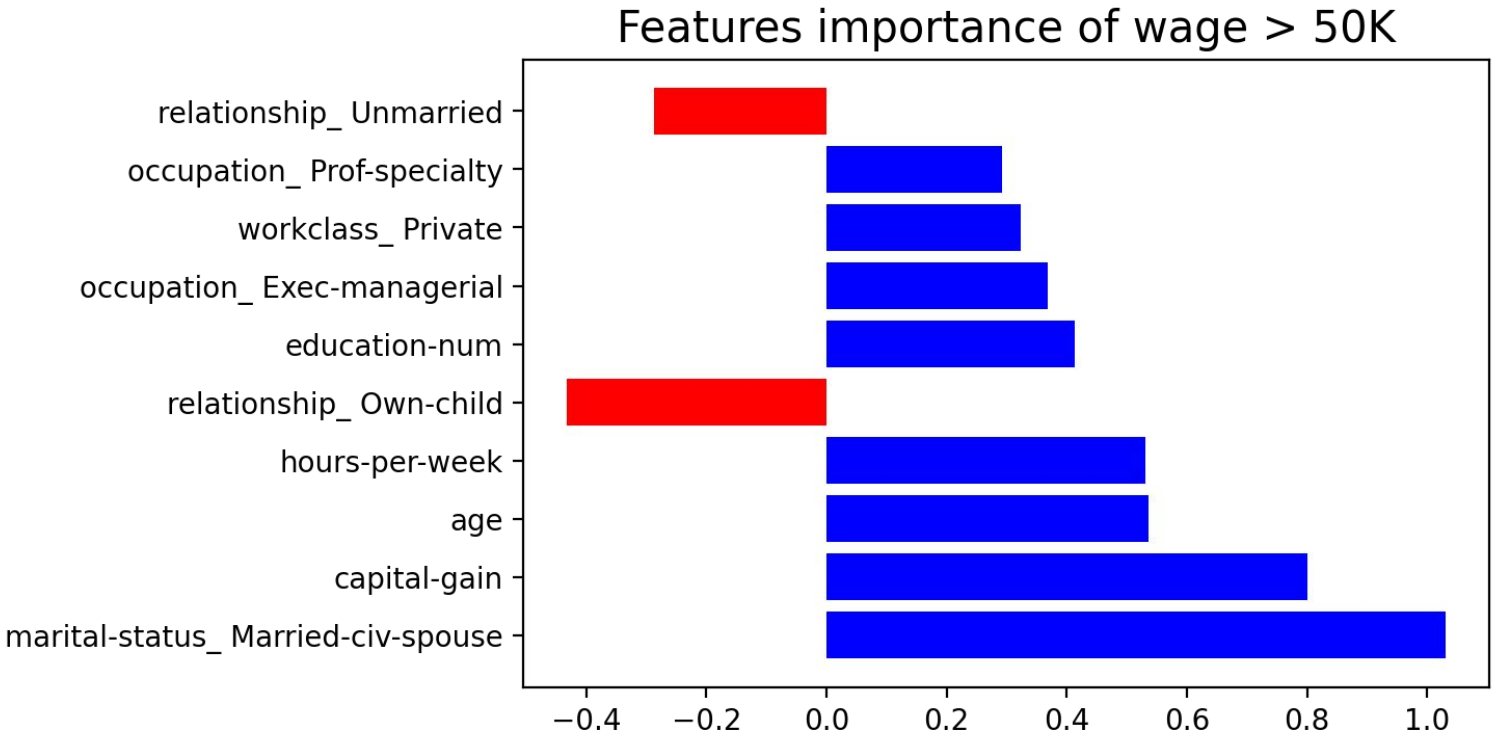
Dataset	Accuracy
Train	0.832
Test	0.81

Train - data

Classes	F1 - score
<= 50 K	0.87
> 50 K	0.67

Real F1-score of test data : 0.797

Features importance from model



- **Age, Country, Work time, Education, Occupation**

What we have tried but did not succeed (for us)

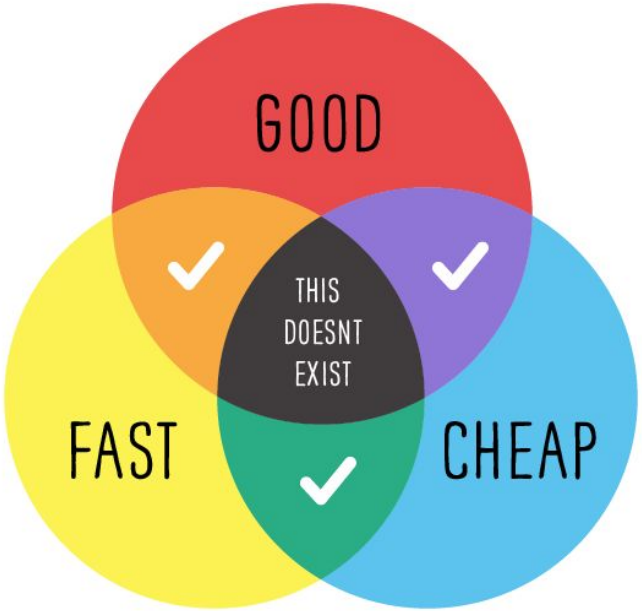
- Regularization - L1
- DecisionTreeClassifier
- RandomForestClassifier
- AdaBoostClassifier
- GradientBoostingClassifier
- Train all data

Conclusion

- Age, Country, Work time, Education, Occupation features effects the wage
- The model using Logistic regression algorithm
- The model got F1-score of test data at 0.797
- The model is not overfit and still beat the baseline score.

Suggestion

- Gain more data
- Use others classifier



Thank You

HEINLEY