Census Income Study

Data Cleaning and Predictive Modeling

Andrew Graham – Fall 2022

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

Overview

Summary and Goals

Purpose

The purpose of this data is to create a prediction model to determine the income of an individual based on given census data. The goal of this data has been binned into having a salary of > 50k and < 50k. The data is from the US census bureau. Logistic Regression, Random Forest and XGBoost modeling will be Evaluated

Overview

Data Overview - Meta

- 2 Files were provided containing data split into training and testing data
- A meta file was provided containing column information, such as: name, description, and types (nominal/continuous)
  - This file contained a few tables which were combined using a fuzzy matching algorithm and then the resulting information allowed the data to be labeled
  - The meta document also contained the unique values from the data which was used to assist in data cleaning efforts
- Instructions were given in the document to drop a column (Instance Weight), so those instructions were followed.
- Data consist of 41features (40 input and 1 output) with 299,285 records

Overview

Data Overview

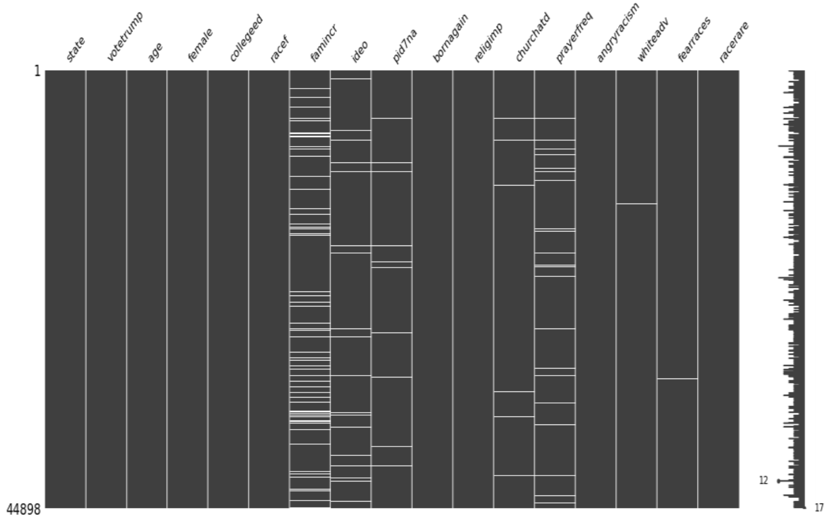
- 40 Input features
  - 7 numerical/continuous
  - 33 nominal
- The 7 continuous feature did not have missing data although a majority had 0 values
- 14 nominal features had complete data
- 19 features were incomplete
  - 14 features were dropped for having over 30% missing data that could not be imputed
- 6% of the data was dropped for having missing values
- Target(Salary) is unbalance with more than 90% in the <50k category

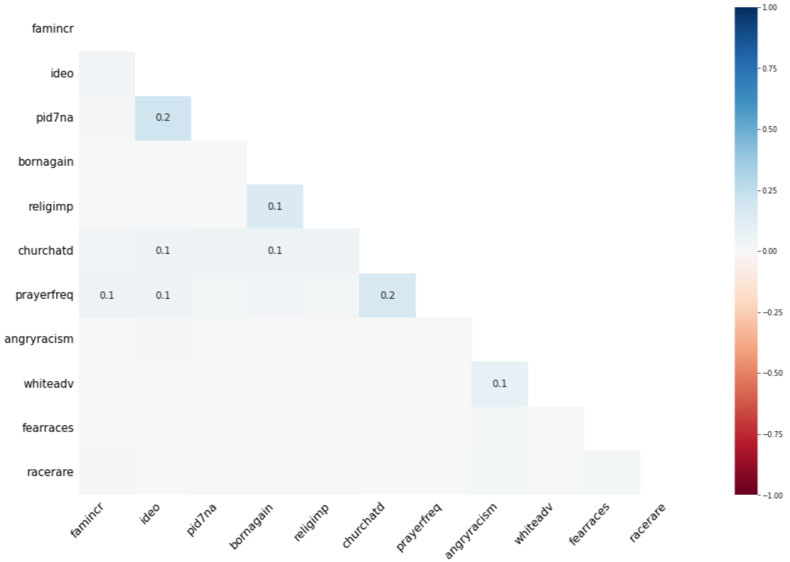
UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

Overview

Data Cleaning

	number_missing	percent_missing
state	0	0.000000
votetrump	0	0.000000
age	0	0.000000
female	0	0.000000
collegeed	0	0.000000
racef	0	0.000000
famincr	4717	10.506036
ideo	1501	3.343133
pid7na	611	1.360862
bornagain	22	0.049000
religimp	20	0.044545
churchatd	322	0.717181
prayerfreq	904	2.013453
angryracism	47	0.104682
whiteadv	52	0.115818
fearraces	100	0.222727
racerare	86	0.191545



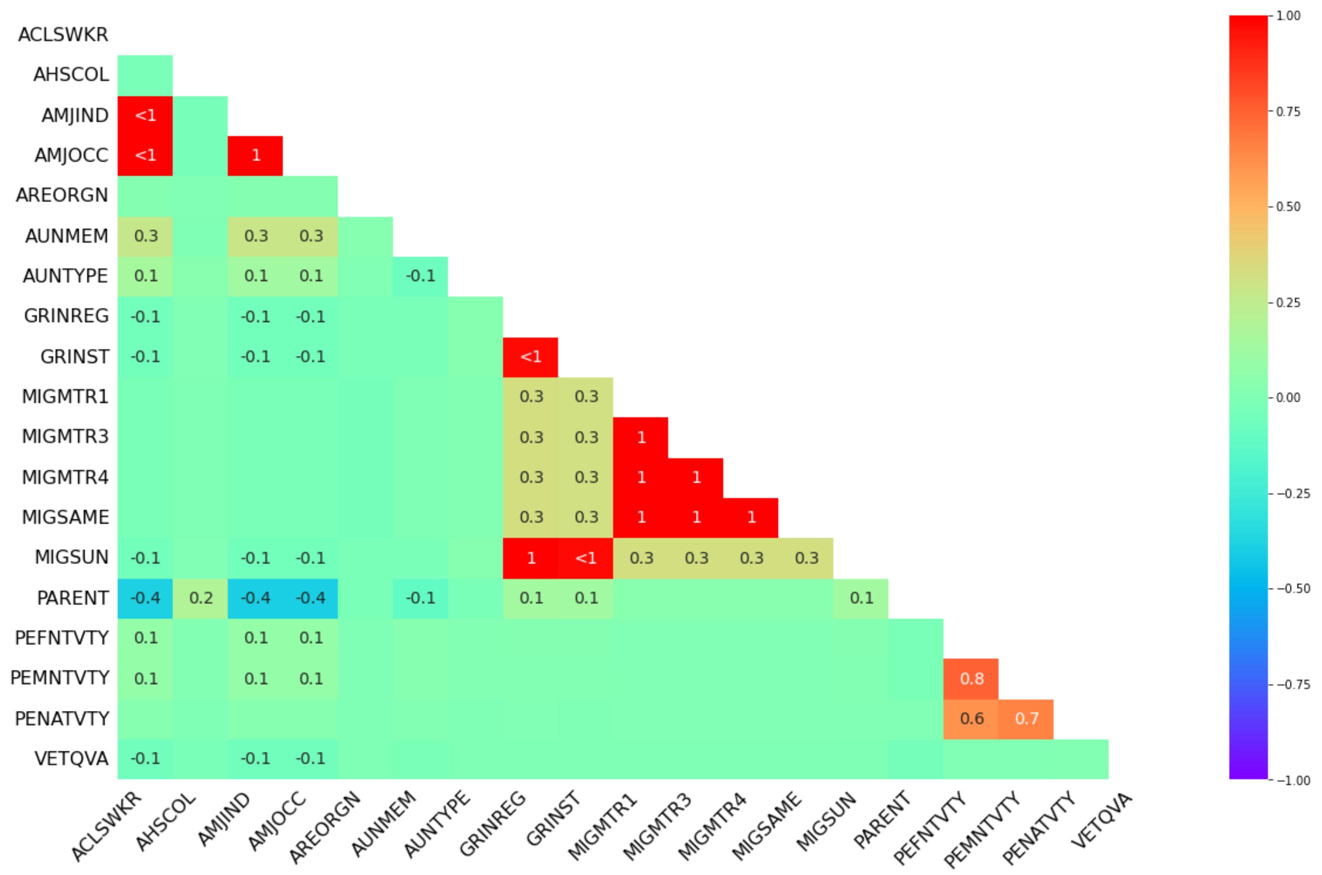


- No misspellings or formatting issues
- No duplicates
- Removed NAs for Target
- 7337 NA
- 16% rows with NAs
  - No dependencies
  - Randomly Distributed between outputs
- NAs Removed
- 37,561 Data Points Remaining

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

EDA

NA Correlation



- From this and the prior chartwe see that most columns look to be MCAR/MAR with the following exceptions...
- -AMJOCC and AMJIND (Major Industry Code and Major Occupation Code)
- This makes sense as they seem to be referencing the same thing
- -GRINREG and GRINST (region and state of previous residence)
- This makes sense as one is dependent of the other.
- -MIGMTR1,MIGMTR3,MIGMTR4 (Migration code Data)
- -PEFNTVTY and PEMNTVTY (Birth pace of Parents)

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

EDA

NA Feature Dropping

Since the missingness looks to be random and using a threshold of 30%. The following features should be dropped:

Feature	# Missing	Missingness
AMJIND	84080	0.362789
ACLSWKR	83508	0.360321
AMJOCC	84080	0.362789
MIGSAME	114346	0.493381
MIGMTR4	114346	0.493381
MIGMTR3	114346	0.493381
MIGMTR1	114346	0.493381
AUNMEM	203225	0.876877
PARENT	203808	0.879392
GRINREG	208751	0.900721
MIGSUN	208751	0.900721
GRINST	209776	0.905143
AHSCOL	215546	0.930040
AUNTYPE	222633	0.960619
VETQVA	228779	0.987138

With the following to be kept:

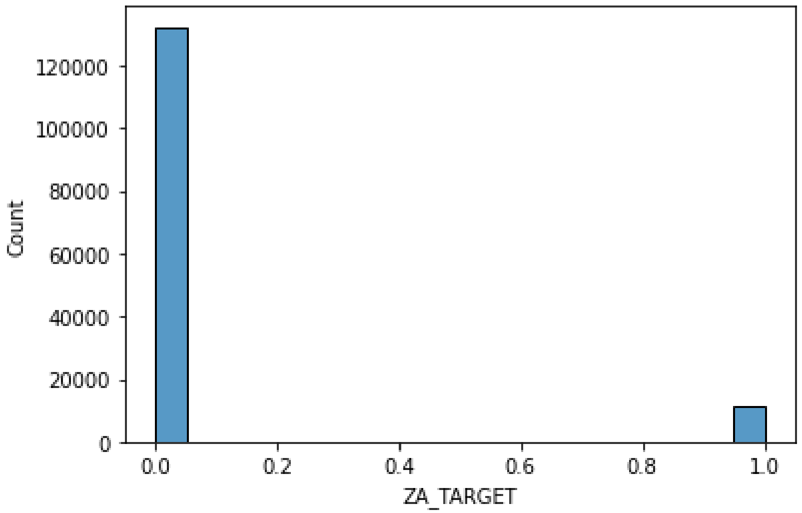
Feature	# Missing	Missingness
AREORGN	1672	0.007214
PENATVTY	5057	0.021820
PEMNTVTY	8779	0.037880
PEFNTVTY	9690	0.041810

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL
OF ENGINEERING & COMPUTER SCIENCE

EDA

Target Variable

- Target Feature is binned at <50k and >50k
- This was converted to 1 for >50 and 0 for <50
- Data was clean and had no missing values



The target variable is highly unbalanced, and this will have to be considered for model creation.

EDA

Numerical Features

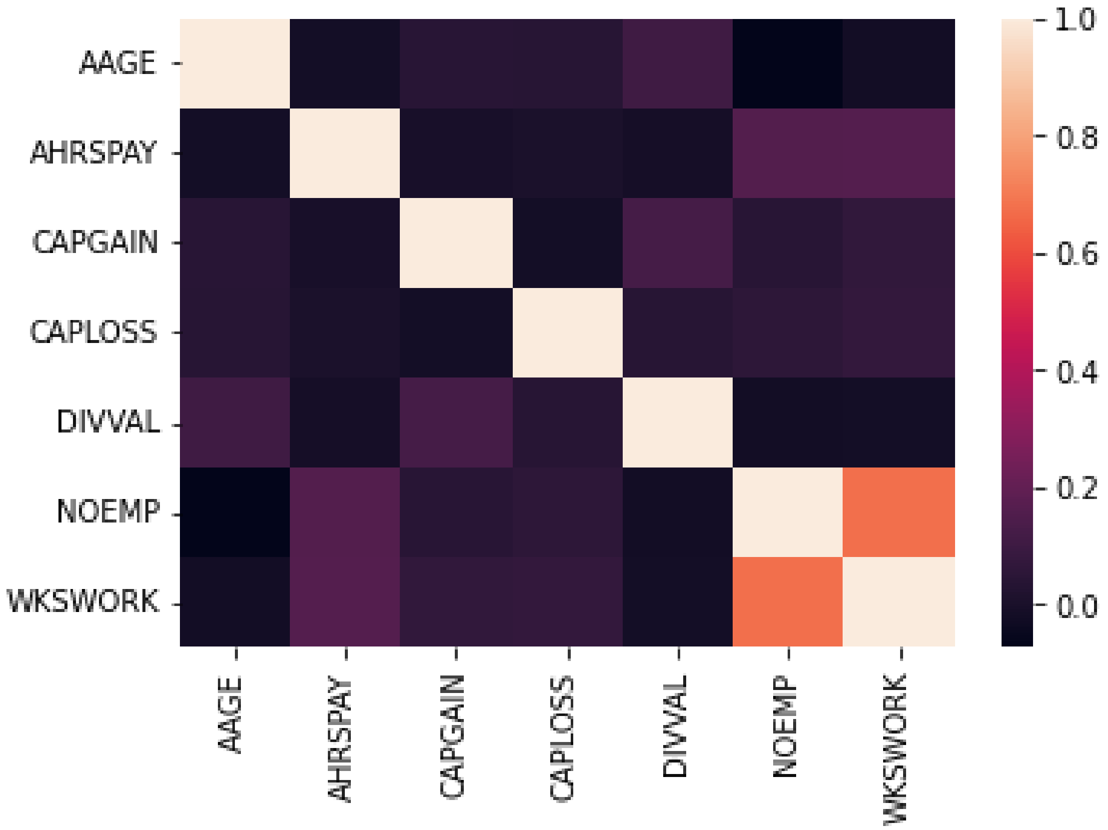
	mean	median	min	max	var	std	skew
AAGE	34.538998	33.0	0.0	90.0	4.981140e+02	22.318468	0.372785
AHRSPAY	55.105027	0.0	0.0	9999.0	7.471515e+04	273.340729	8.878780
CAPGAIN	431.742176	0.0	0.0	99999.0	2.181608e+07	4670.768536	19.090569
CAPLOSS	36.849010	0.0	0.0	4608.0	7.278652e+04	269.789771	7.685924
DIVVAL	195.851259	0.0	0.0	99999.0	3.755251e+06	1937.847082	27.144287
NOEMP	1.956172	1.0	0.0	6.0	5.592548e+00	2.364857	0.752317
WKSWORK	23.178375	8.0	0.0	52.0	5.955560e+02	24.404016	0.210018

- 7 continuous numerical features.
- AHRSPAY (Wage per hour), CAPGAIN, CAPLOSS, and DIVAL are all highly right skewed.
- AHRSPAY, CAPGAIN and DIVAL all have ceiling of 9999

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

EDA

Numerical Features



- Low correlations between most of the variables
- NOEMP and WKSWORK have high correlation
  - This is understandable as someone who has employees likely works a higher amount of weeks

EDA

Nominal Features Summary

- Most features can be left as is
- The features with larger number of categories can have them consolidated as many categories only show <50k
- Country of organ for mother, father, and self may consider changing to USA not USA if performance looks to be ian issue
- Year could possibly be deleted
- Education should be updated to Ordinal

Model Evaluation

Test Train Split

```
train = df[df['ZZ_SPLIT'] == 'train'].reset_index()
X_train = train.drop(columns=['ZA_TARGET','ZZ_SPLIT'])
y_train = train['ZA_TARGET']


test = df[df['ZZ_SPLIT'] == 'test'].reset_index()
X_test = test.drop(columns=['ZA_TARGET','ZZ_SPLIT'])
y_test = test['ZA_TARGET']
```

- Train Test split 80:20
- Split was given by files
- Data was converted into ordinal, numerical, and nominal features converted into 1 hot encoding
- Data scaled using MinMax Scaler

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

Model Evaluation

Logistic Regression, Random Forest, XGBoost

```
logR = LogisticRegression(max_iter= 10000)

param_grid_logR = {'C' : [0.001,0.01,0.1,1,10,100]}


grid_logR = GridSearchCV(estimator=logR, param_grid=param_grid_logR,cv=8)
grid_logR = grid_logR.fit(X_train,y_train)
```

```
----- Logistic Regression ----
Accuracy : 0.9203716314816617
f1 : 0.1982788777723361
Testing
Accuracy : 0.9222181823078685
f1 : 0.06928778601353529
```

```
rf = RandomForestClassifier(n_jobs=14)

param_grid_rf = {'max_depth': [1,2,4,8,16],
                'n_estimators':[100,150,200],
                'max_features':['sqrt',1,5,10,20],
                'min_samples_split': [1,2,4,6]}


grid_rf = GridSearchCV(estimator=rf, param_grid=param_grid_rf,scoring='None',cv=8)
grid_rf = grid_rf.fit(X_train,y_train)
```

```
----- Random Forest ----
Accuracy : 0.9305310410962712
f1 : 0.31145395688408295
Testing
Accuracy : 0.9318466448511291
f1 : 0.2971809470906819
```

```
xg = xgb.XGBClassifier()

param_grid_xg = {"gamma": [0.1,0.2,0.3,0.4,0.5],
                "max_depth": [1,2,3,4,5], # default 3
                "n_estimators": [100,150,200]}


grid_xg = GridSearchCV(estimator=xg, param_grid=param_grid_xg,cv=8)
grid_xg = grid_xg.fit(X_train,y_train)
```

```
----- XGBoost ----
Accuracy : 0.9342904395925118
f1 : 0.4463700234192038
Testing
Accuracy : 0.9365060127391966
f1 : 0.4482153306026916
```

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL
OF ENGINEERING & COMPUTER SCIENCE

Important Features

Logistic Regression

Logistic Regression

	Feature	Coef	Coef_abs	Coef_odds	Coef_prob
3	AHGA	4.293125	4.293125	73.194849	0.986522
0	AAGE	2.461090	2.461090	11.717576	0.921369
12	WKSWORK	1.863244	1.863244	6.444612	0.865675
5	CAPGAIN	1.133288	1.133288	3.105851	0.756445
6	CAPLOSS	1.068971	1.068971	2.912380	0.744401

XGBoost

	Feature	Coef	Coef_abs	Coef_odds	Coef_prob
3	AHGA	4.59835	4.59835	99.320302	0.990032
0	AAGE	2.73568	2.73568	15.420226	0.939099
12	WKSWORK	1.90952	1.90952	6.749848	0.870965
5	CAPGAIN	1.14659	1.14659	3.147442	0.758888
6	CAPLOSS	1.09426	1.09426	2.986972	0.749183

- Most Important Feature for both Models: AHGA (Education)
- Logistic Regression
  - If Education increases by 1 odds of an income >50k increase be 73
- XGBoost
  - If Education increases by 1 odds of an income >50k increase be 99

Conclusion

- XGBoost Performed the best on the test set with 93.6% accuracy
- Test and Train set Accuracy were dimilar, so did not suffer from Over/Under fitting
- F1 score low, so mostly guessed with the majority result (income<50k)
- XGBoost best model to move forward