Marketing based on income classification

Jorge Duran
Dec 2022

Introduction

Problem/Challenge:

- Businesses struggle to be effective in their marketing to the right customer segments. If businesses can predict a customer's income based on certain features, they can target products and services more effectively.

Objectives:

- Understand what features can predict if a customer's income is greater than $50k USD based on US Census data.
- Build a predictive classification model with 85% precision.

Dataset

- Consolidated data from the 1994 US Census - UCI ML Repository

- It consists of 48,842 records and 14 features(mix of continuous and discrete)

- Divided in 3 files:
  - “adult.data” with 32,561 records (train set)
  - “adult.test” with 16,282 records (test set)
  - “adult.names” containing information on the columns names

- Dataset will be used to predict if a person’s yearly income is greater than $50k

Data Wrangling

- No missing values, only “?” values in the following features:
  - workclass: 2799 records with “?” (~5.7%)
  - occupation: 2809 records with “?” (~5.7%)
  - native-country: 857 records with “?” (~1.7%)

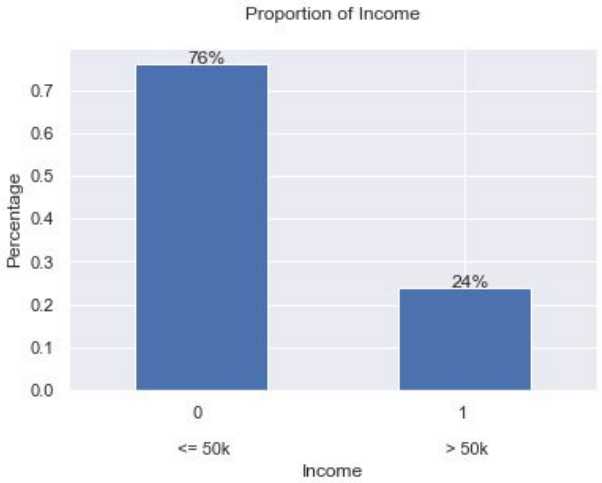
- Since the features with “?” values were categorical, we changed them to “unknown” so they have their own category

- Feature “fnlwgt” was removed - it’s a field internally calculated by the Population Division at the Census Bureau and it’s not clear how this field is calculated and what it means

- 2 more features created by bucketing the “age” and “hours-per-week” features

Exploratory Data Analysis

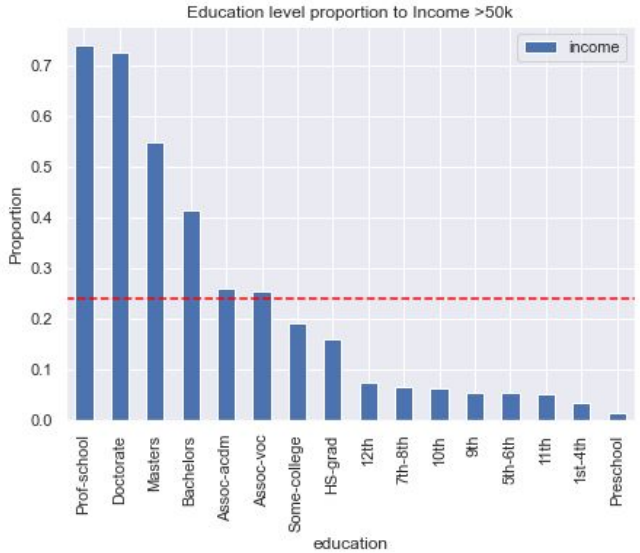
The distribution of our target variable “income” is shown below.



Some of the main factors that could contribute to an individual’s income can be education, industry, location, work type, age and gender.

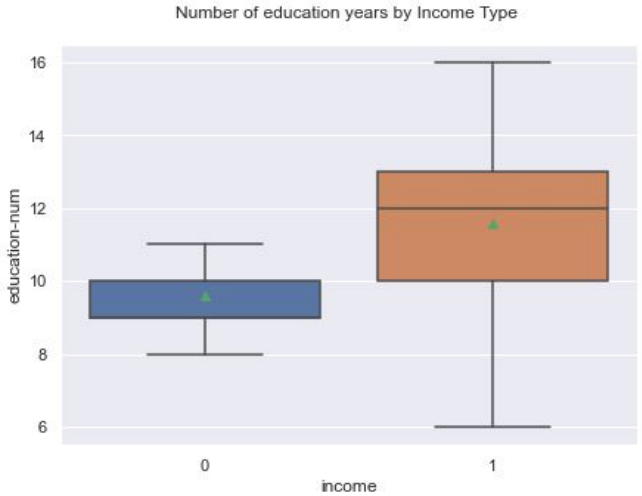
Education level

Professional school, Bachelors, Masters and Doctorate had the highest proportion of income greater than $50k.
Professional school ranks above the other 3 and in this context corresponds to what it seems an education level higher than Masters.



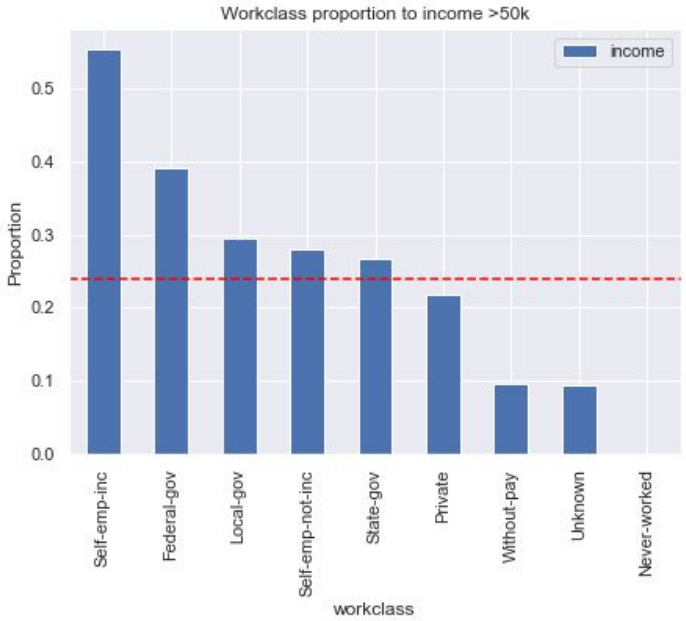
Years of Education

About 75% of people with income greater than $50k have at least 10 years of education, which corresponds to at least some college.



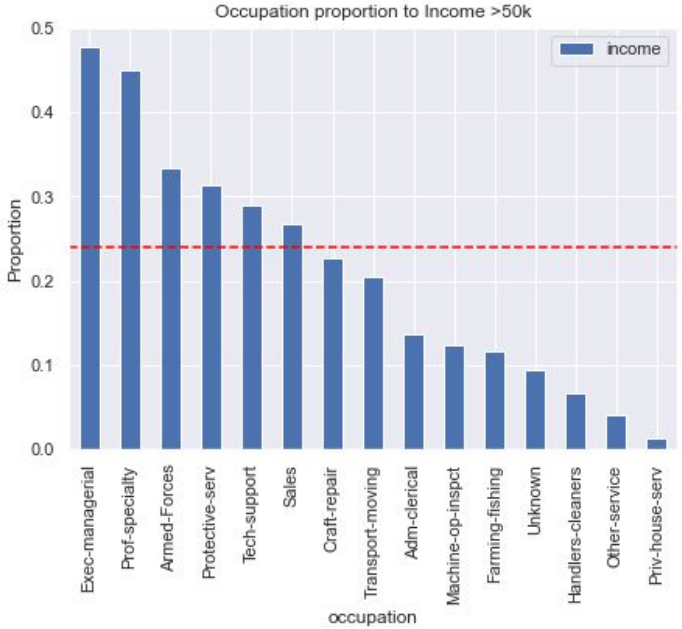
Workclass

Self-emp-inc had the highest proportion of higher income, but this category only accounts for 3.5% of our data. The private sector accounts for almost 70% of our data and it appears to lean more towards lower income.



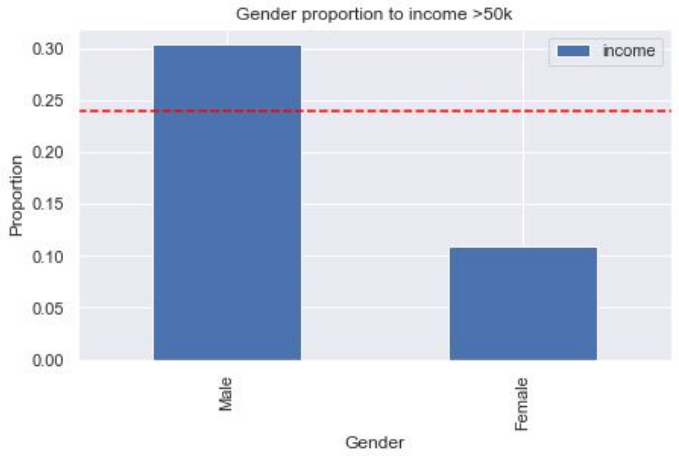
Occupation

Exec-managerial and Prof-specialty are the occupations with a higher proportion of high income and they account for 25% of our data.



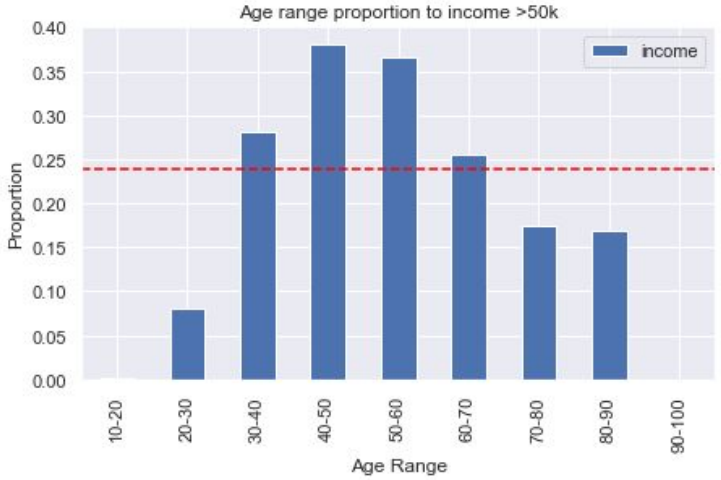
Gender

There is a strong differentiation in income due to gender. Being male almost triples the probability of higher income over being female. Our data also show 67% of the census respondents were males and 23% were females.



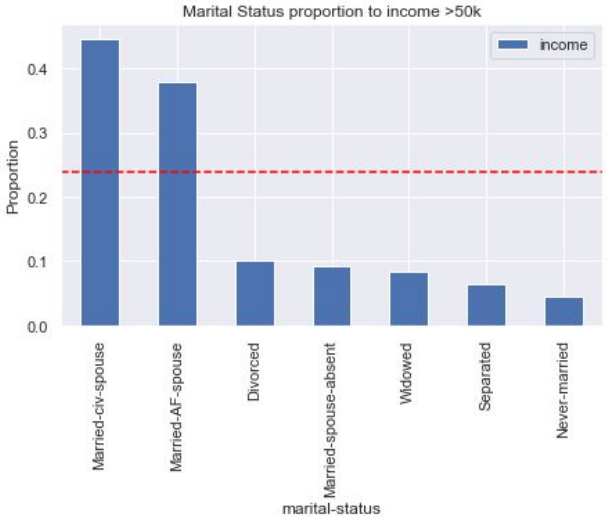
Age

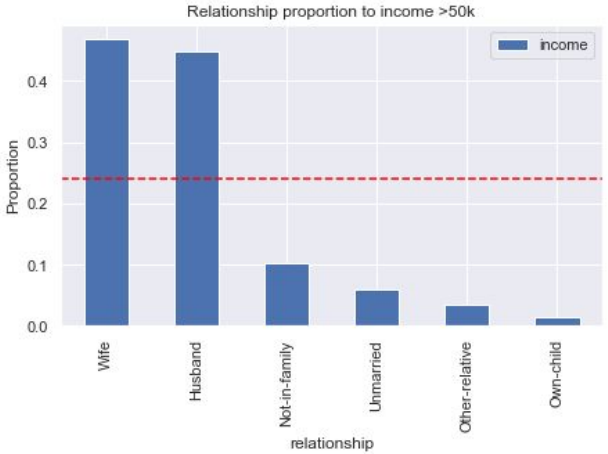
Age ranges 40 to 60 show higher proportion of high income. We then see a decline after 60 years, which could be related to people's retirement.



Marital Status / Relationship

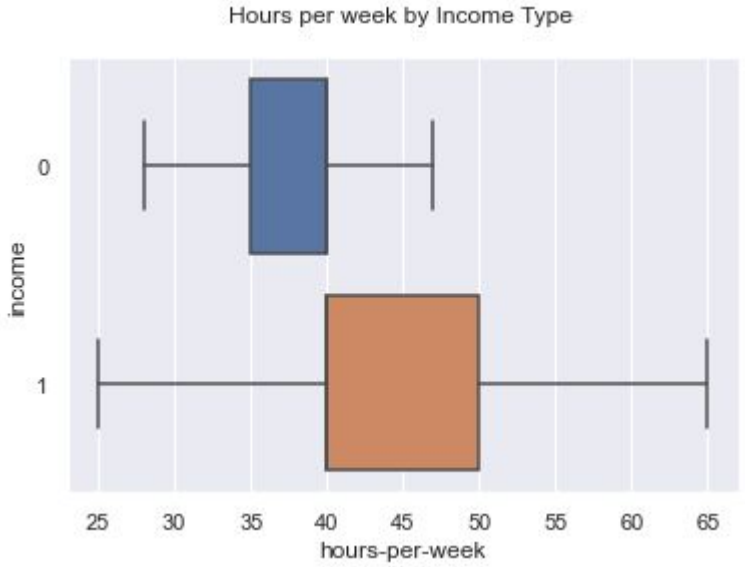
There is a better probability of high income if you're married. The size of the difference is somewhat surprising, especially the almost >3x jump in probability of having high income if you are married vs if you're not.





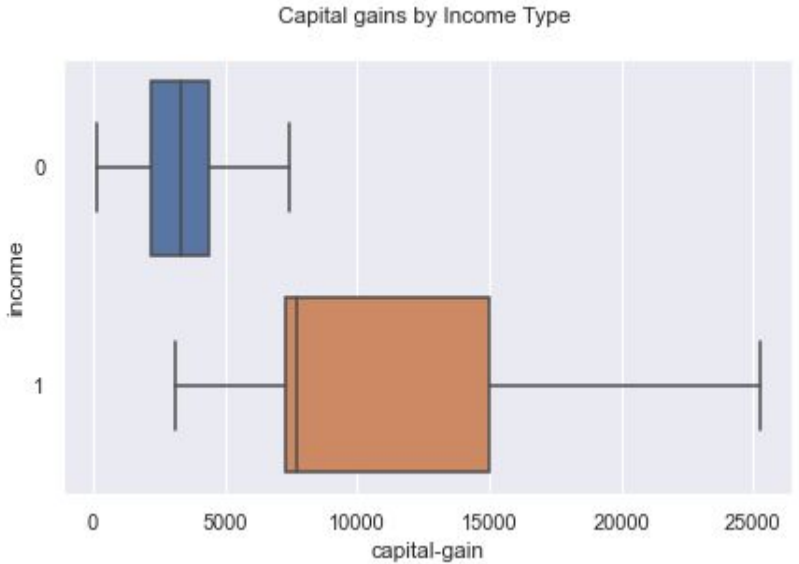
Hours worked per week

40 hours or more have a higher probability of high income. 75% of the high income earners worked 40 or more hours.



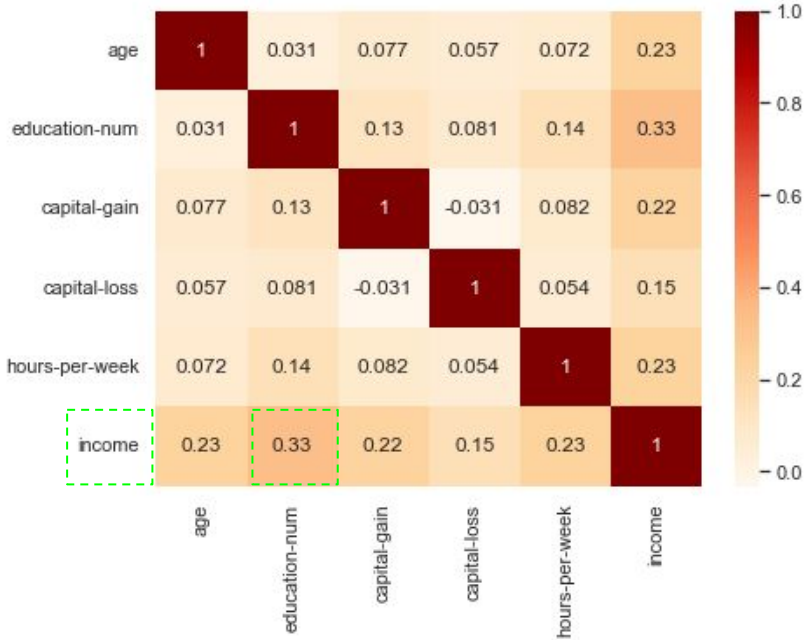
Capital gains

Majority of people with high income have capital gains of $7,000 or more. Although this is a clear indicator, only 9% of people in our dataset have capital gains.



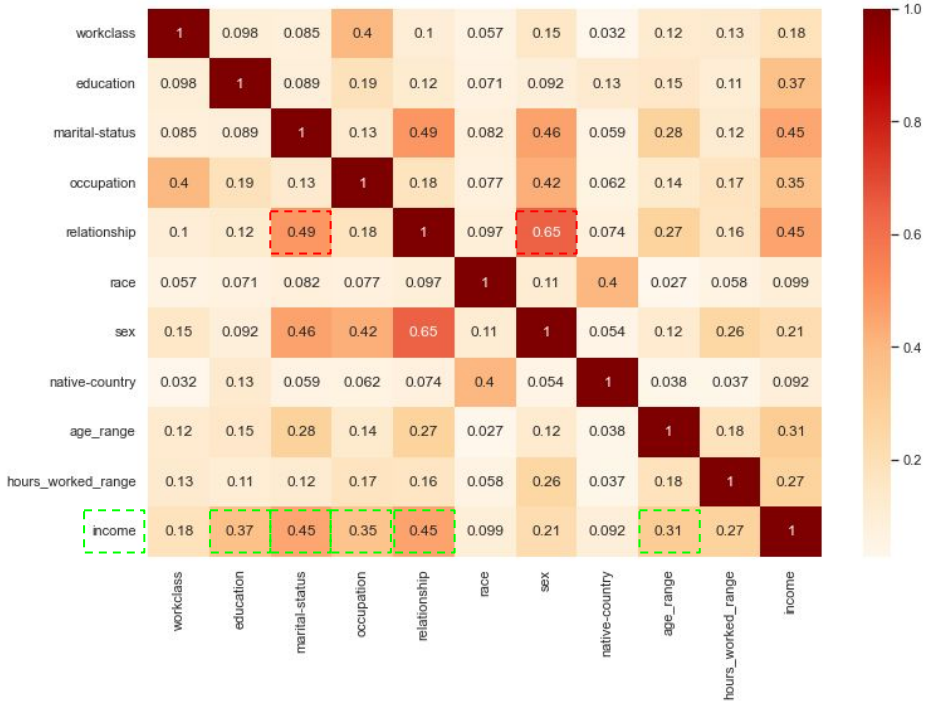
Correlation heatmap (continuous variables)

Some factors we found to influence higher income are level of education, age, occupation, marital status and gender. We also see that there is little multicollinearity amongst the continuous features that we need to worry about.



Correlation heatmap (categorical variables)

Marital status and relationship show a moderate correlation, which it's expected since both features identify whether the person is married or not.
Also there is a moderate correlation between sex and relationship, this was unexpected. We can say that depending on your gender, this determines the role you play if you're married (husband or wife).



Feature Selection / Preprocessing

- For categorical features we grouped values with a low contribution to reduce the number of features after one-hot encoding.

- Encoded categorical features using one-hot encoding and dropping the first value. Dataframe went from 16 to 61 features after encoding.

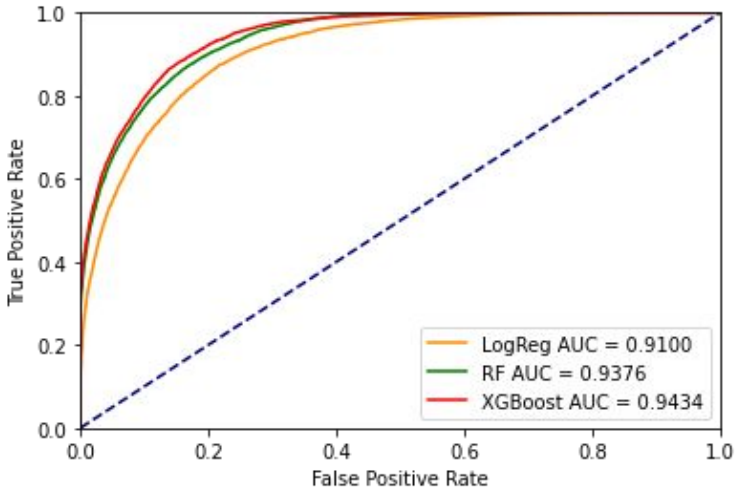
- Scaled numerical features using Robust Scaler and Standard Scaler.

- Train/test split (70/30). The split distribution retained the same ratio as the original dataframe.

Modeling

- 3 models were built for evaluation: Logistic Regression, Random Forest & XGBoost
- Metric used to assess performance was the ROC AUC
- For hyperparameter tuning we used Random Search with 5-fold cross validation

Model Name	Best Hyperparameters	ROC AUC
Logistic Regression	solver = 'saga' penalty = 'elasticnet' c = 10 l1_ratio = 0.5 max_iter = 1200	0.9100
Random Forest	n_estimators = 700 min_samples_split = 5 min_samples_leaf = 9 max_features = 0.4 max_depth = 30	0.9376
XGBoost	n_estimators = 150 booster = 'gbtree' colsample_bytree = 0.6 max_depth = 5 eta = 0.2 scale_pos_weight = 3.2	0.9434



Top 10 feature importances for the XGBoost model

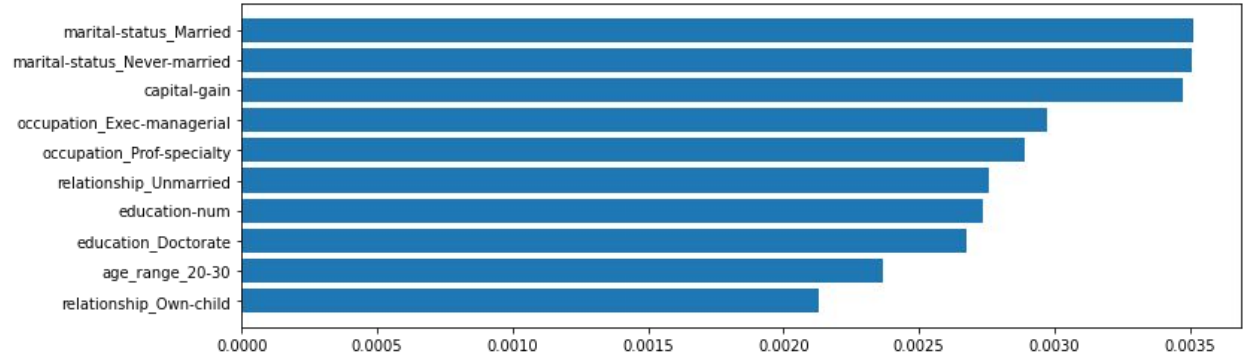
Marital Status

Capital Gains

Occupation

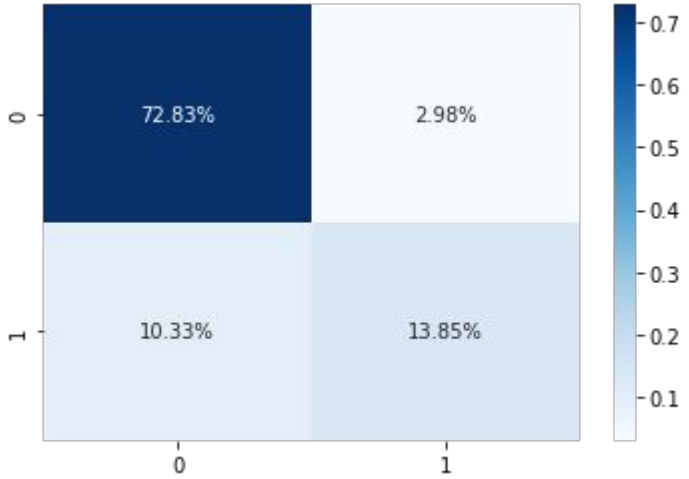
Relationship

Education



Since the predictive model's target was 85% precision, we adjusted our threshold on predictions to 0.8 instead of the default (0.5). (Precision: 85%, Recall: 61% for our training set).

A final validation with the 0.8 threshold was done on the test set to obtain final metrics.



Confusion Matrix

	precision	recall	f1-score	support
0	0.88	0.96	0.92	11109
1	0.82	0.57	0.68	3544
accuracy			0.87	14653
macro avg	0.85	0.77	0.80	14653
weighted avg	0.86	0.87	0.86	14653

Classification Report

Conclusion

- Marital status was the best predictor on high income. It's worth mentioning this only applied to married or unmarried, this didn't matter as much for other cases.

- Other factors that also were good predictors were capital gains, occupation, relationship, education and gender.

- Model can provide business value by helping businesses and marketing teams to infer a customer's income based on these features.

- For future work, the income threshold can be adjusted, more data can be included (i.e. industry, location) and the analysis/models can be done with multiple income brackets to target audiences more effectively.