Marketing based on income classification

Prepared by: Jorge Duran

Dec 2022

Introduction

Businesses want to be effective in their marketing campaign efforts to maximize value by targeting the right customer segment, instead of the general audience. But sometimes they struggle to find effective ways to narrow those specific customer segments. When this is the case they might overspend on marketing efforts that don't yield the expected results or revenue.

One important piece of information in determining a customer's segment is income. If businesses can predict a customer's income based on other pieces of known information, they can target certain products and services more effectively to that customer and it would be more likely for them to pay for those goods or services.

Our company can help those businesses by providing specific customer bases, in this case customers who have an income greater than $50k USD. These could be customers that are in a good position financially and are willing to pay more for better products or services.

The intention of this analysis is to better understand what characteristics can predict whether a customer's income is greater than $50k USD based on US Census data.
With the information obtained in this analysis, our target is to build a predictive classification model to segment people with income greater than $50k USD with a 85% precision.

Dataset

The dataset for this project was obtained from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/adult) and it consists of consolidated data from the 1994 US Census.

The dataset consists of 48,842 records and 14 attributes (mix of continuous and discrete).
Dataset is 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

This dataset will be used to predict if a person's yearly income is greater than $50,000.

Attributes / Features:

- **income**: >50K, <=50K. (*Target variable*)
- **age**: continuous.
- **workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt**: continuous.

- **education**: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- **education-num**: continuous.
- **marital-status**: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- **occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- **relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **sex**: Female, Male.
- **capital-gain**: continuous.
- **capital-loss**: continuous.
- **hours-per-week**: continuous.
- **native-country**: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Data Wrangling

At first glance there were no missing values, all variables had the same number of records. After reviewing more thoroughly we found 3 categorical features with “?” values:

- workclass: 2799 records with “?” (~5.7%)
- occupation: 2809 records with “?” (~5.7%)
- native-country: 857 records with “?” (~1.7%)

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

After reviewing the description of the feature “fnlwgt”, it was decided to remove it as this is 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.

Additionally we created 2 more features by bucketing the “age” and “hours-per-week” features. Age was split in bins of 10, ranging from 10-100 to split age groups. Hours were split in 7 ranges.

Exploratory Data Analysis

The distribution of our target variable “income” is shown below on Fig 1. 76% are people with income less or equal than $50k, the remaining 24% have income greater than $50k.

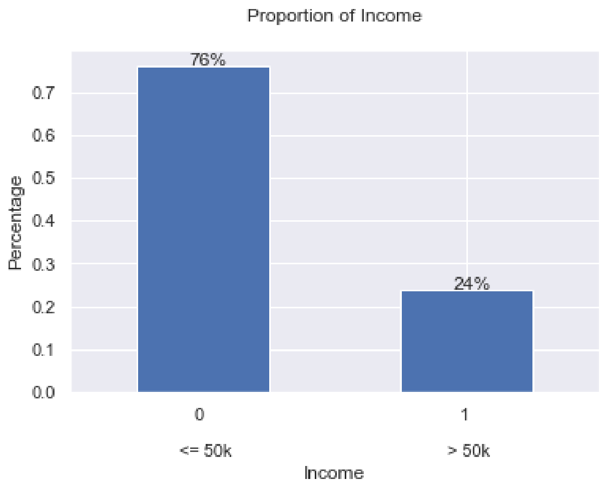
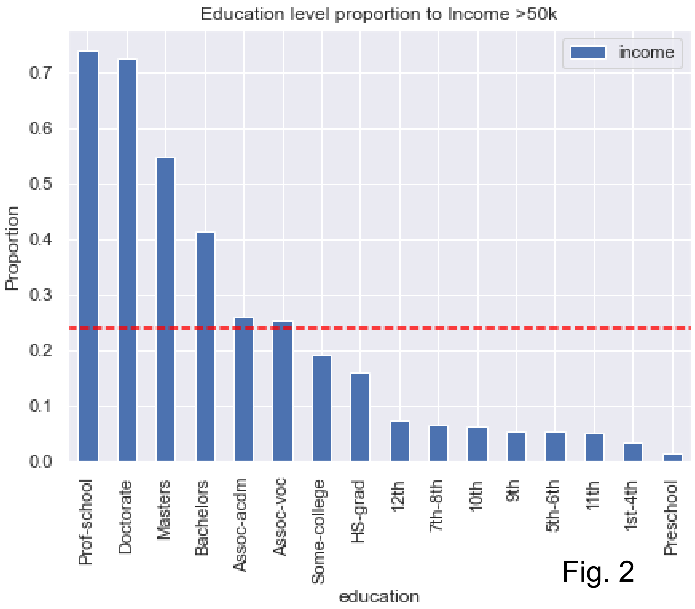
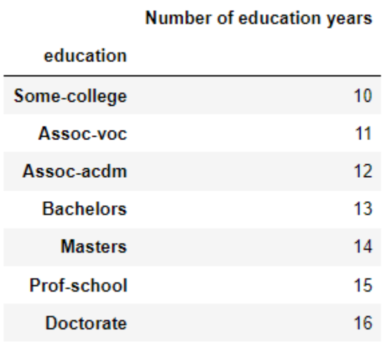


Fig. 1

Some of the main factors that could contribute to an individual’s income can be level of education, industry, location, work type, age and gender. We explored the ones included in our dataset.

Usually a high level of education could result in a job that is well paid. Our data confirmed our assumption, Fig. 2 shows Professional school, Bachelors, Masters and Doctorate as having the highest proportion of income greater than $50k ($p << 0.01$). It’s worth mentioning that Professional school ranks above the other 3 and in this context corresponds to what it seems an education level higher than Masters.

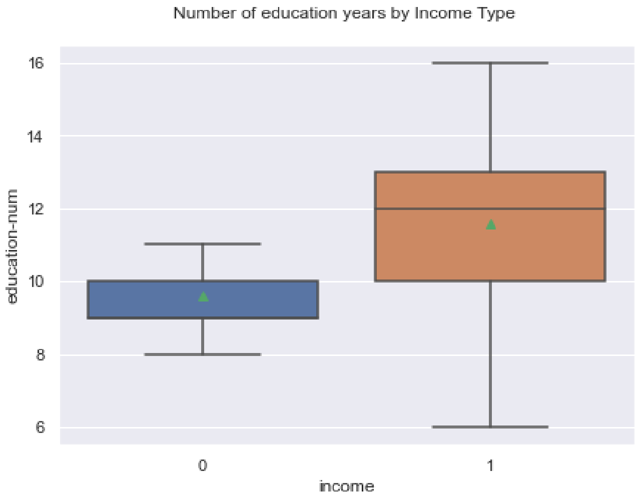


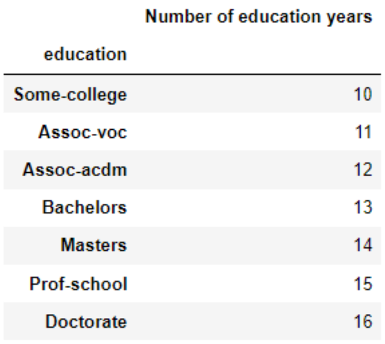


education	Number of education years
Some-college	10
Assoc-voc	11
Assoc-acdm	12
Bachelors	13
Masters	14
Prof-school	15
Doctorate	16

Fig. 2

We also can take a look at the number of years of education on Fig 3. About 75% of people with income greater than $50k have at least 10 years of education (p << 0.01), which corresponds to at least some college.



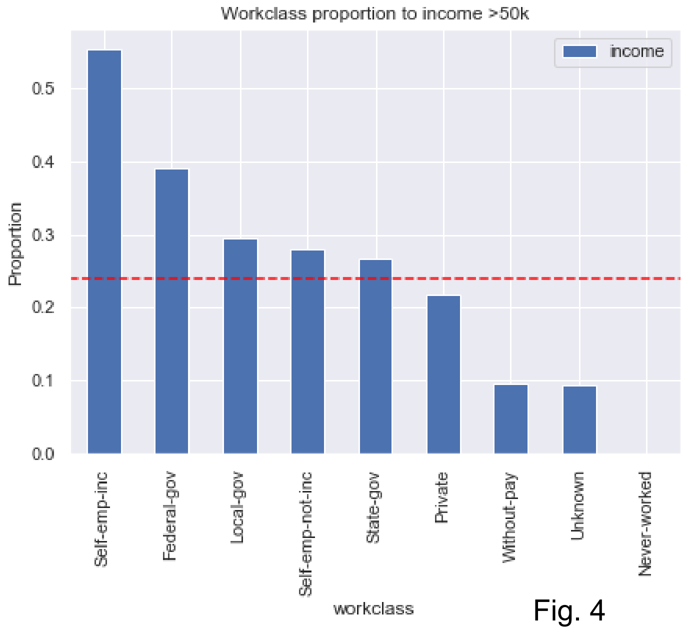


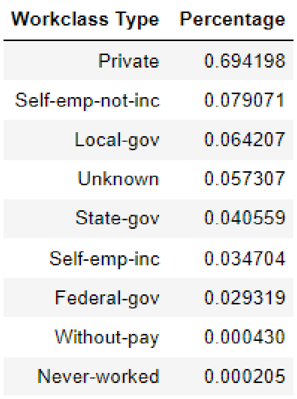
education	Number of education years
Some-college	10
Assoc-voc	11
Assoc-acdm	12
Bachelors	13
Masters	14
Prof-school	15
Doctorate	16

Fig. 3

Other factors we had mentioned at the beginning are industry, location and work type. Unfortunately we don’t have location nor industry data. Regarding work type we can look at workclass and occupation.

Fig. 4 shows us workclass distribution. Self employed (Inc) seems to have the highest proportion of higher income (p << 0.01), 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 the income being less than $50k (p << 0.01).

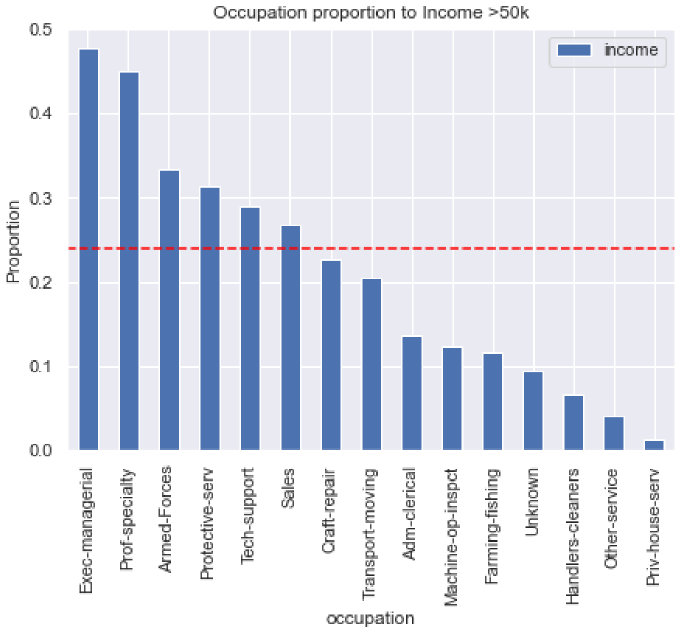


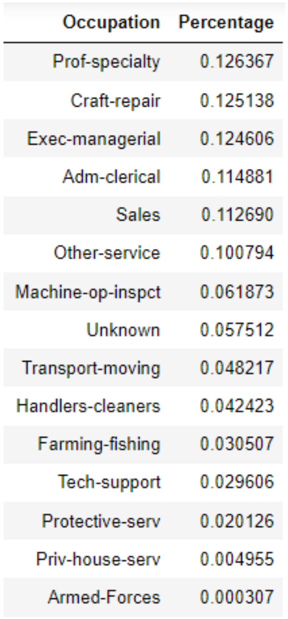


Workclass Type	Percentage
Private	0.694198
Self-emp-not-inc	0.079071
Local-gov	0.064207
Unknown	0.057307
State-gov	0.040559
Self-emp-inc	0.034704
Federal-gov	0.029319
Without-pay	0.000430
Never-worked	0.000205

Fig. 4

Following our exploration of work type, Fig 5 shows us the occupation breakdown. As expected, Exec-managerial and Prof-specialty are the occupations with a higher proportion of income greater than $50k ($p << 0.01$) and they account for 25% of our data.



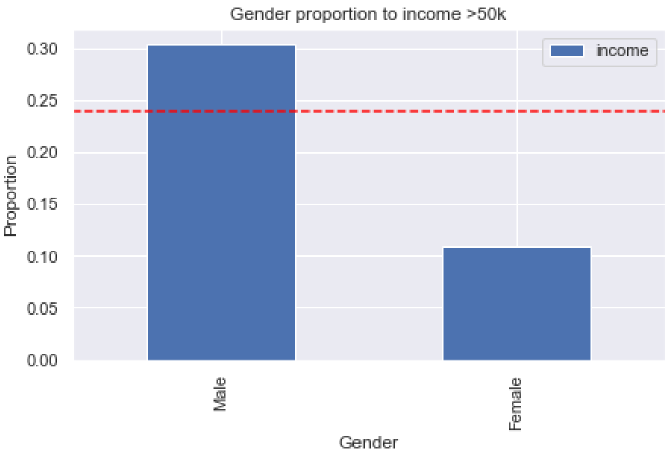


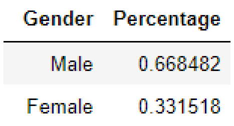
Occupation	Percentage
Prof-specialty	0.126367
Craft-repair	0.125138
Exec-managerial	0.124606
Adm-clerical	0.114881
Sales	0.112690
Other-service	0.100794
Machine-op-inspct	0.061873
Unknown	0.057512
Transport-moving	0.048217
Handlers-cleaners	0.042423
Farming-fishing	0.030507
Tech-support	0.029606
Protective-serv	0.020126
Priv-house-serv	0.004955
Armed-Forces	0.000307

Fig. 5

Let’s explore our remaining factors from the initial list: gender and age.

Fig 6 shows a strong differentiation in income due to gender. Being male almost triples the probability of higher income over being female ($p << 0.01$). These data from 1994 still represent a more favorable income if the individual is a male. Our data also show 67% of the census respondents were males and 23% were females.

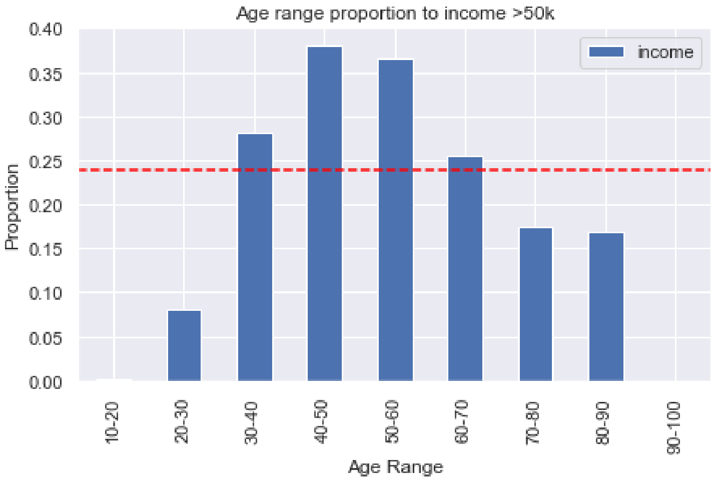


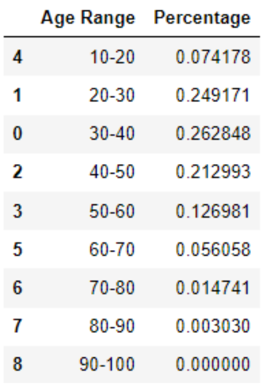


Gender	Percentage
Male	0.668482
Female	0.331518

Fig. 6

Finally let’s explore if age is relevant for higher income. Fig. 7 shows higher proportions between ages 40 to 60 ($p << 0.01$). We then see a decline after 60 years, which could be related to people’s retirement ($p >> 0.05$), but this showed not to be statistically significant.





	Age Range	Percentage
4	10-20	0.074178
1	20-30	0.249171
0	30-40	0.262848
2	40-50	0.212993
3	50-60	0.126981
5	60-70	0.056058
6	70-80	0.014741
7	80-90	0.003030
8	90-100	0.000000

Fig. 7

During our data analysis, we found other factors that could contribute to high income: marital status/relationship, the amount of hours worked per week and capital gains from investments.

Both Fig. 8 and Fig. 9 show that you have a better probability of high income if you’re married ($p << 0.01$). The size of the difference here is somewhat surprising, especially the almost >3x jump in probability of having income >50k if you are married vs if you are divorced. While it seems likely that most people’s jobs wouldn’t change over the course of a divorce, one possible explanation could be that income issues are a large factor in many divorces. This would likely require an entire separate study to understand better.

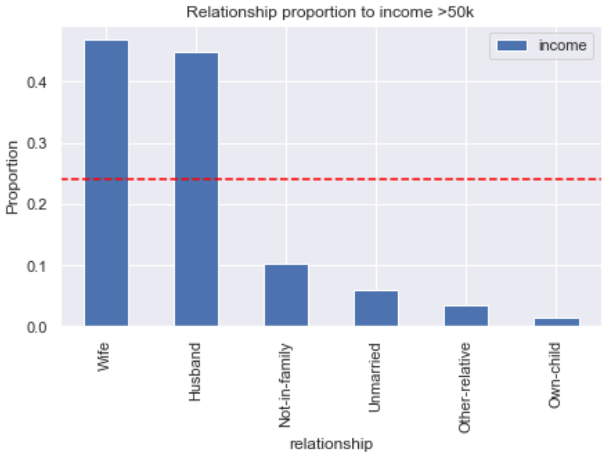


Fig. 8

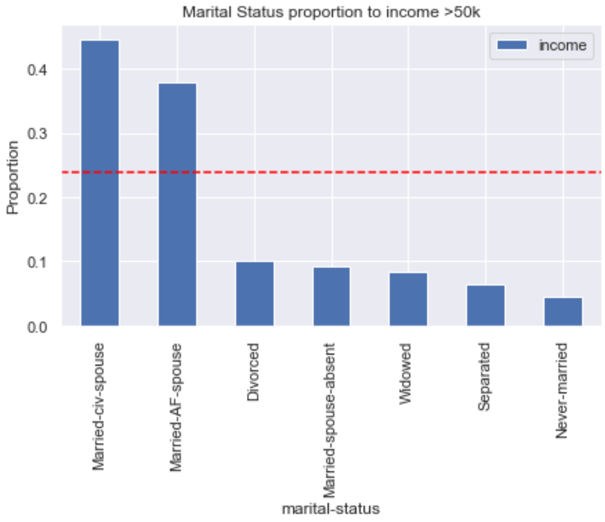


Fig. 9

Fig. 10 shows that 40 hours or more have a higher probability of high income (p << 0.01). Fig. 11 shows different ranges of hours worked per week. According to an article by Chris Kolmar, earnings rise as a function of hours worked up to the 55 hours per week mark, we find this to be true for our dataset.

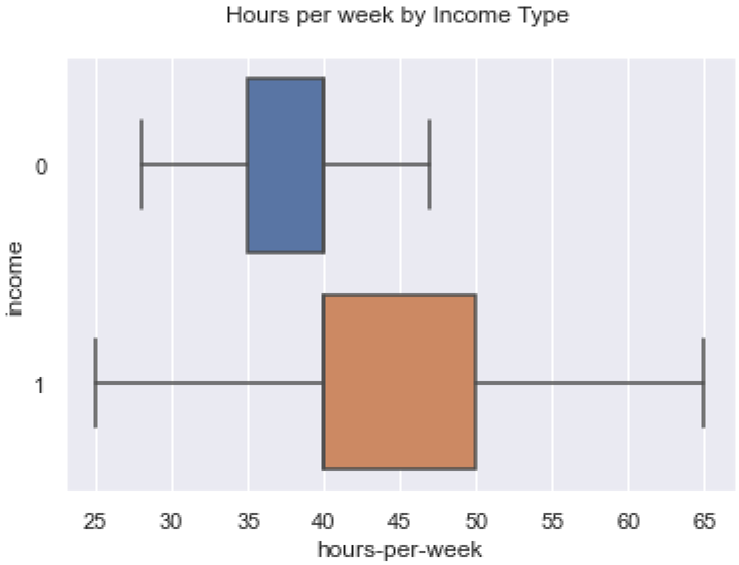
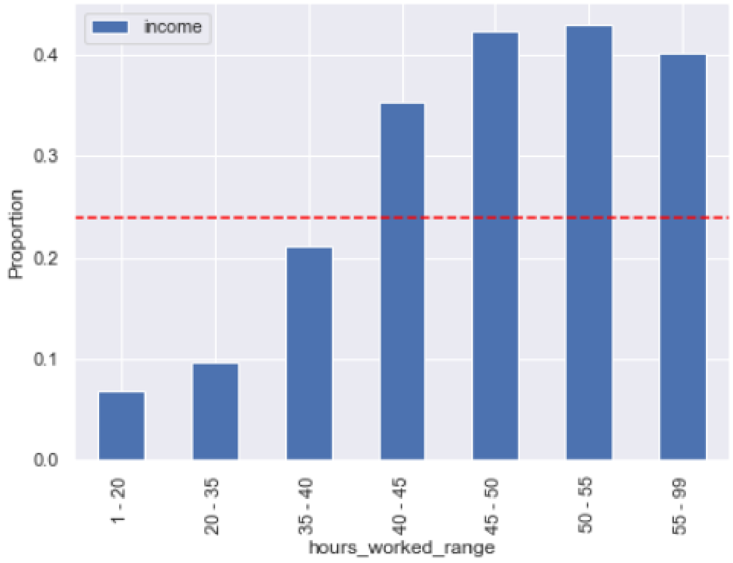
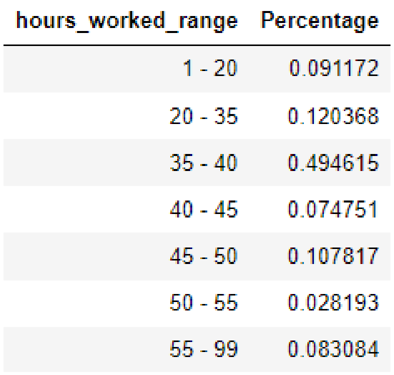


Fig. 10





hours_worked_range	Percentage
1 - 20	0.091172
20 - 35	0.120368
35 - 40	0.494615
40 - 45	0.074751
45 - 50	0.107817
50 - 55	0.028193
55 - 99	0.083084

Fig. 11

Capital gains chart on Fig. 12 shows that the majority of people with incomes greater than $50k have capital gains of $7,000 or more ($p << 0.01$). Although this is a clear indicator, only 9% of people in our data have capital gains.

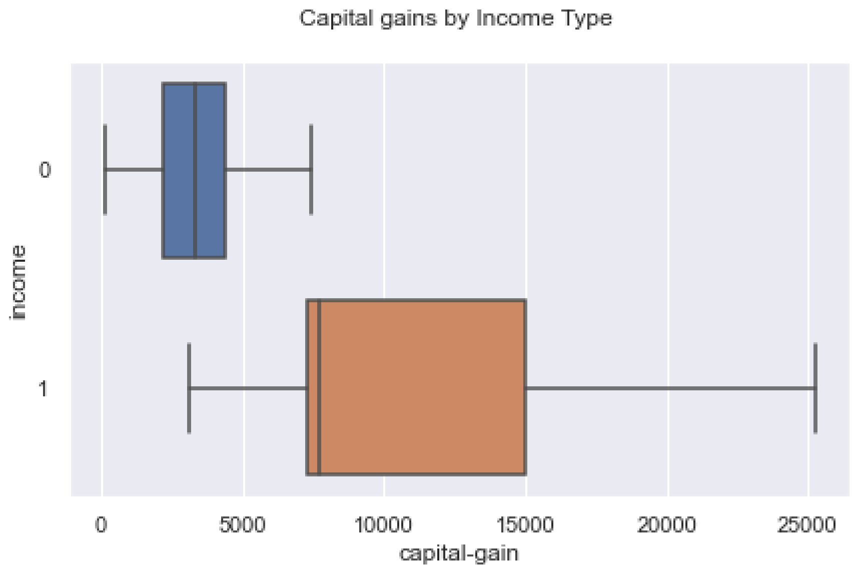


Fig. 12

The following correlation heatmap in Fig. 13 confirms some of the factors we found to influence higher income, such as 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.

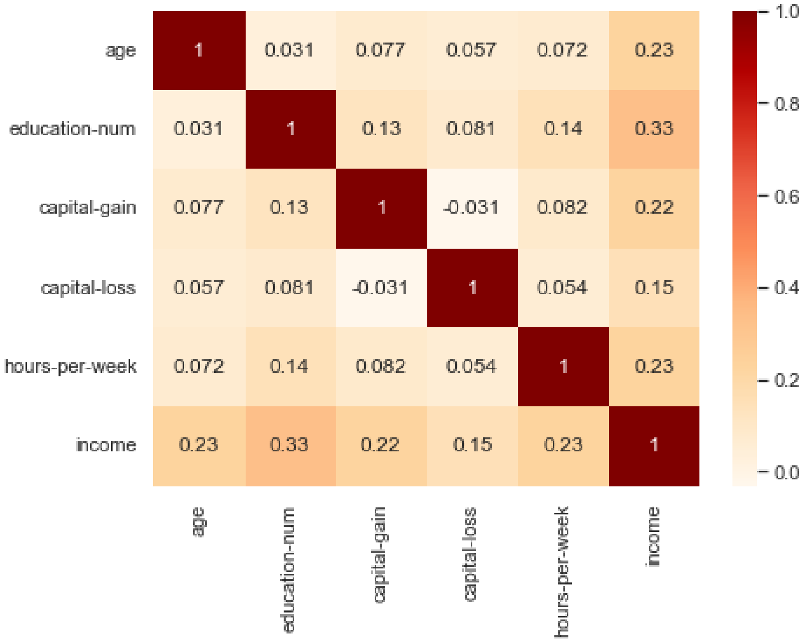


Fig. 13

Fig. 14 looks at categorical variables using Cramer's V to approximate an appropriate correlation number. We can observe that 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. After further exploration we can say that depending on your gender, this determines the role you play if you're married (husband or wife).

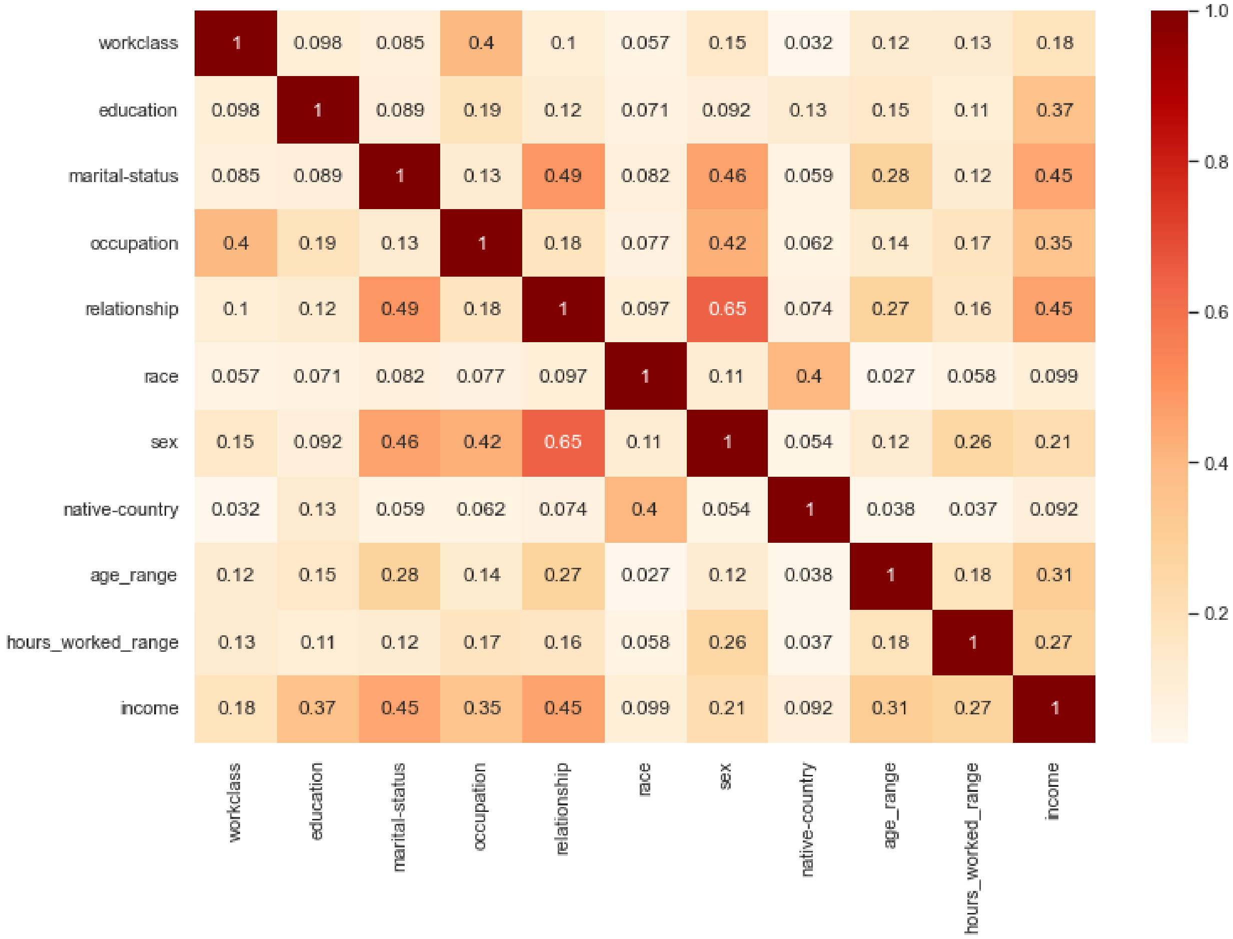


Fig. 14

Finally we built a basic Random Forest model to obtain the features importances, with the following results on Fig. 15:

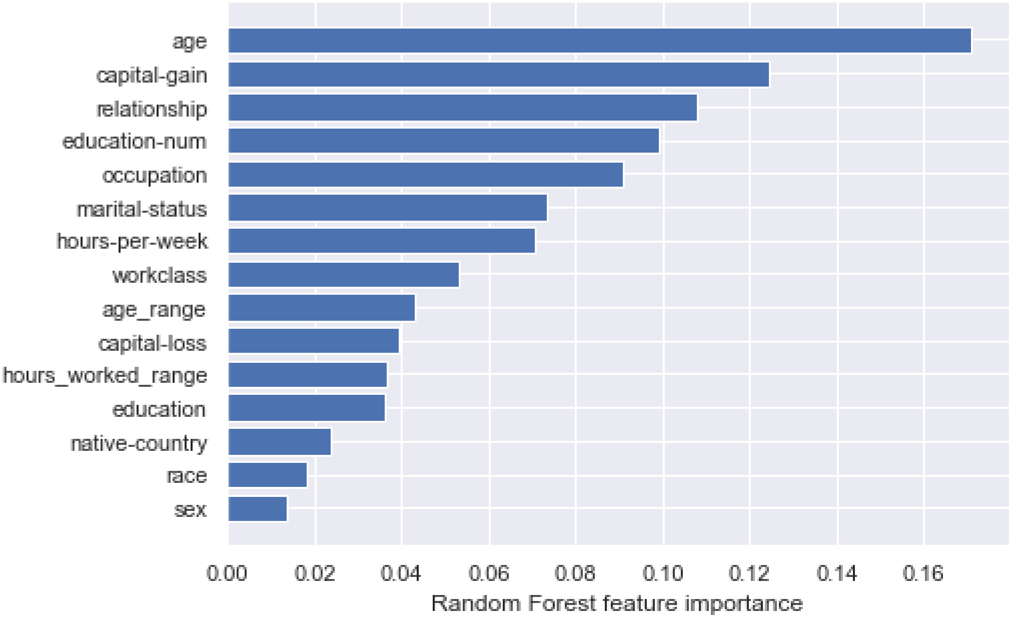


Fig. 15

Again we see many of the same factors we had discussed before as having an influence on income, with age and capital gains coming out on top. However, the model puts sex and education with lower importance than expected. This could be explained by the model potentially getting most of its information from other variables such as the top ones shown in the chart.

Feature Selection / Preprocessing

For our categorical features we grouped values with a low contribution to that feature as “Other” category to reduce the number of features once we do one-hot encoding. This was done based on 2 conditions:

- A percentage threshold - any categories with less than 1% will be grouped as 'Other' category
- Exclude categories from grouping if that category was identified on the EDA as having value on the prediction (example: people with Doctorates represent a small percentage, but it's a good indicator on predicting income)

workclass: grouped 'Without-pay' and 'Never-worked' as “Other”
education: grouped all grades below 12th as a single category “12th and under”
marital status: grouped 'Married-civ-spouse', 'Married-spouse-absent' and 'Married-AF-spouse' as “Married

occupation: grouped 'Armed-Forces' and 'Priv-house-serv' as “Other”
relationship: grouped 'Husband' and 'Wife' as “Other
native-country: grouped other than US as “non-US”, left unknown as their own category
age_range: grouped '70-80', '80-90', '90-100' as “Older than 70”

After this grouping, we proceeded to encode our 10 categorical features using one-hot encoding and dropping the first value. Our dataframe went from 16 to 61 features after encoding. We removed age and hours-per-week since we created new categorical features based on those.
To scale numerical features we used Robust Scaler ‘education-num’ and Standard Scaler for capital gains and losses.
Finally we split our dataset in 70% training and 30% testing. These resulted in 34,189 records for training and 14,653 records for testing. The split distribution seems to be in the same ratio as the original dataframe (24% with income > 50k, 76% with income <=50k).

Modeling

For initial model selection, 3 models were built: Logistic Regression, Random Forest & XGBoost. The metric used to assess performance was the ROC AUC, the thinking being that it gives a threshold-independent idea of the predictive power of the models across many different business scenarios. For hyperparameter tuning we used mainly Random Search with 5-fold cross validation.

Results and tuned hyperparameters are shown in Fig. 16 and Fig. 17.

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

Fig. 16

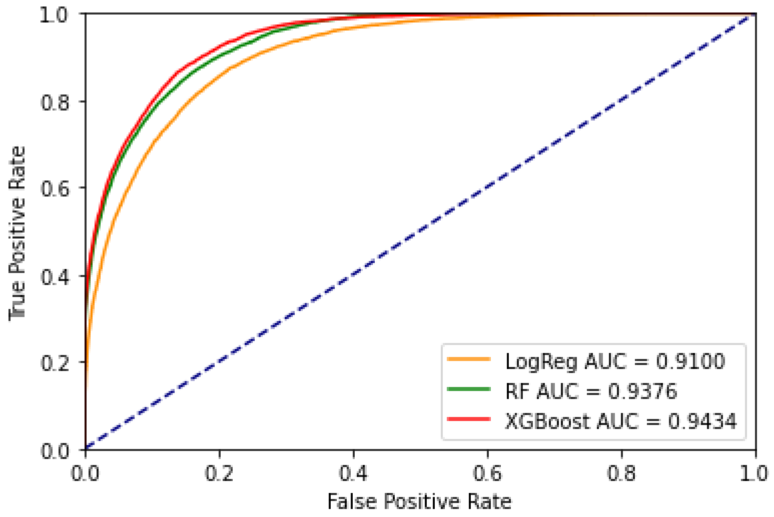


Fig. 17

XGBoost was the model with the best performance (AUC = 0.9434). Fig. 18 shows the Top 10 feature importances from this model.

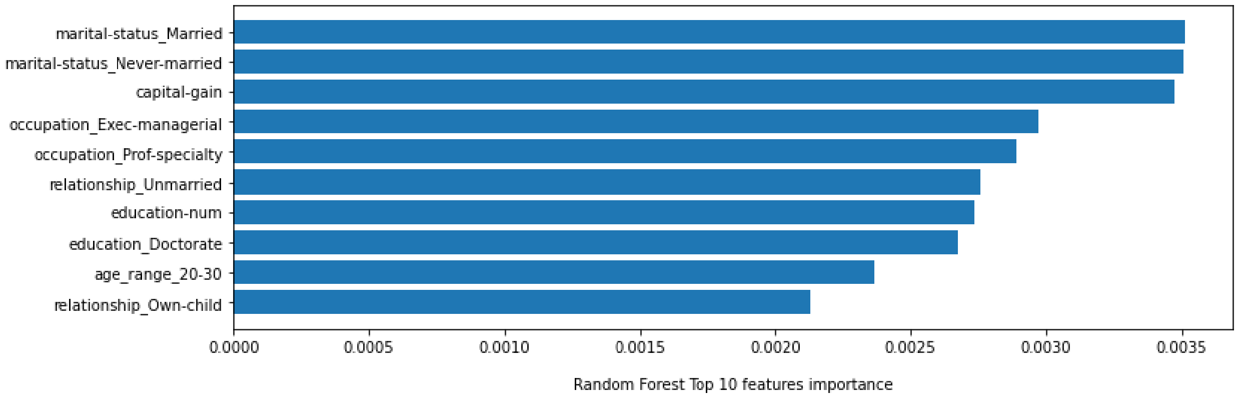


Fig. 18

We then proceeded to adjust the threshold based on our objective to obtain a precision of 85%. The reason for the 85% precision target is because we want the model to minimize false positive predictions. This way we avoid sending unnecessary marketing emails to customers that don’t have an income greater than $50k. If we start spamming customers with income below $50k, we risk them changing their attitude towards the brand and thinking the company’s products are getting more expensive.

Fig. 19 shows precision, recall and F1 score for different thresholds. In this case we found 0.8 to have a precision of 85% and recall of 61% for our training set.

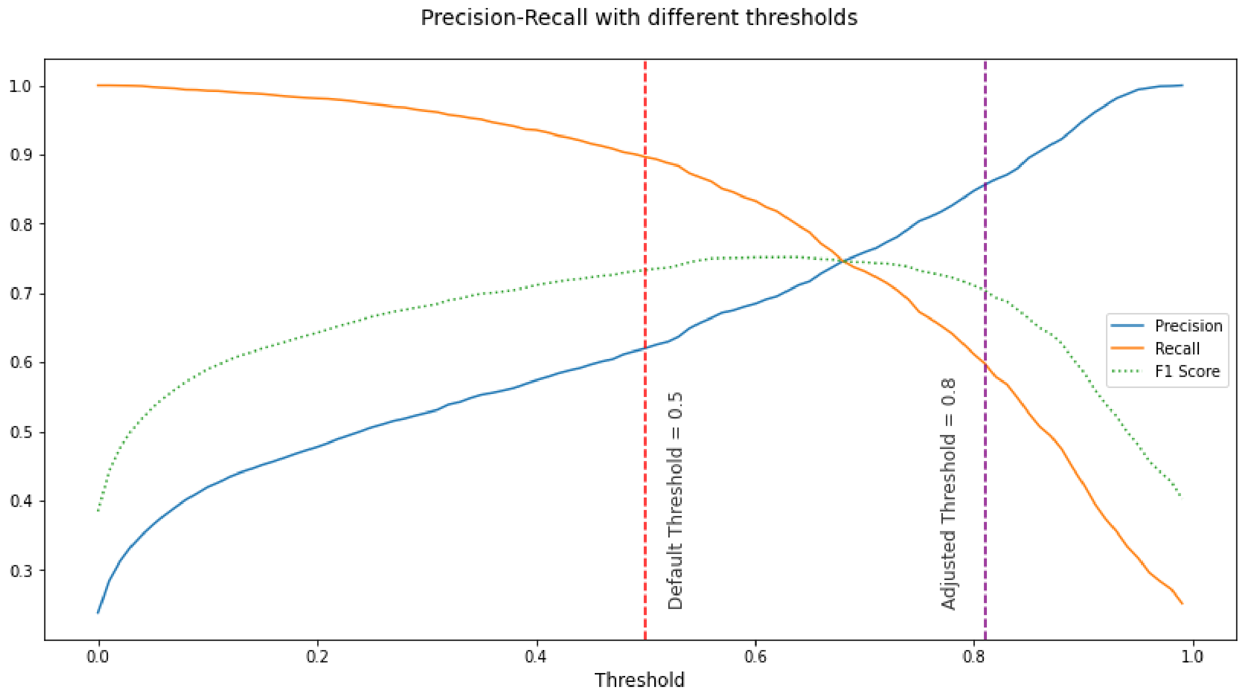


Fig. 19

A final validation with the 0.8 threshold was done on the test set to obtain final metrics. Fig. 20 shows the confusion matrix and Fig. 21 shows the classification report.

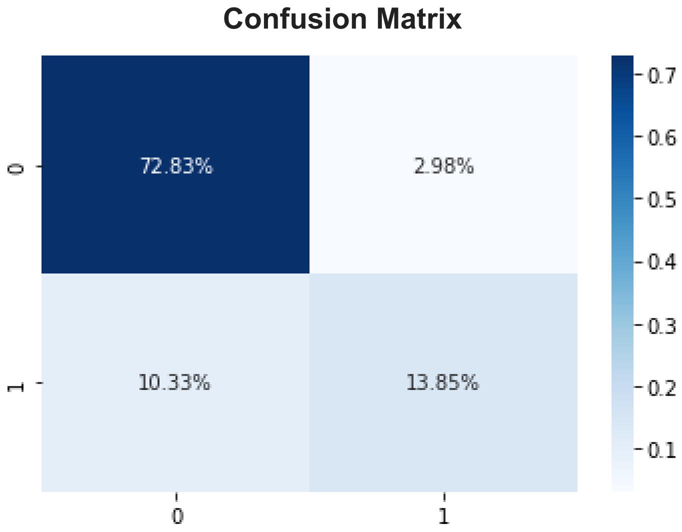
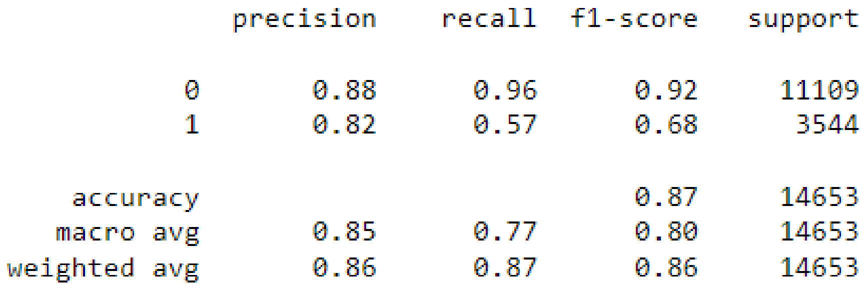


Fig. 20

Classification Report



```
              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
```

Fig. 21

Precision for our test set went a bit lower (82%) than our training set (85%). Similar for recall, it dropped 4 points. Model doesn't look to be overfitted and it generalizes well.

Conclusion

As expected, our analysis showed that some categories such as level of education, occupation and gender are good predictors on a customer’s income. But it was surprising to find that marital status was the best predictor on whether income is greater than $50k. It’s worth mentioning this only applied to married or unmarried, this didn’t matter as much for other cases like divorced, separated or widowed. Another factors that also were good predictors were capital gains and losses, the assumption can be made that people with better incomes are in a better position financially to make investments.

Our final predictive model had a precision of 83%, 2 points below our initial target of 85% but seems to generalize well to unseen data. As such, it does seem that this model would be able to provide business value by helping businesses and marketing teams to infer a customer’s income based on other factors, and thus create more effective and robust customer segments for their marketing efforts.

Although the analysis and models were completed with 1994 US Census data, the process can be adjusted for newer or more relevant data. Most likely it will result in some variation due to considerable changes in the work space, such as improvements on gender equality and new occupations (i.e. big tech).

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