US Census People Analytics Case Study Report

Abhishek Kumar

Business Problem

Creating a machine learning model to predict whether an individual's income is >50K or <=50K per year from the census data.

My assumption:

I have assumed that this data is to be used by a startup credit lending company and they want to classify individuals into low risk and high risk categories for credit card approval. My main aim with the final model would be to reduce misclassification rate of individuals who have earnings '<=50K' into '>50K'

Preliminary Analysis

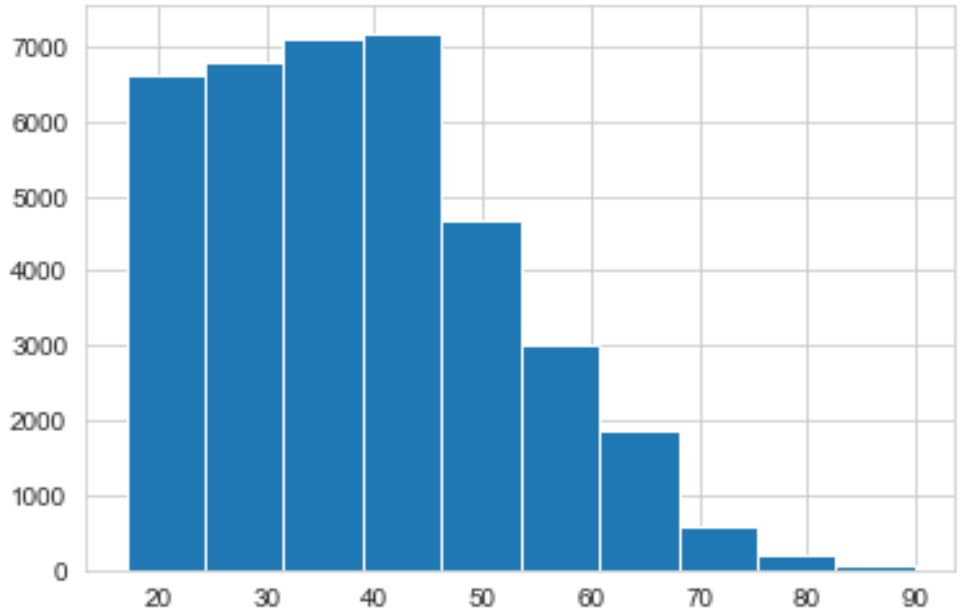
1. The dataset had whitespaces in almost all the categorical columns

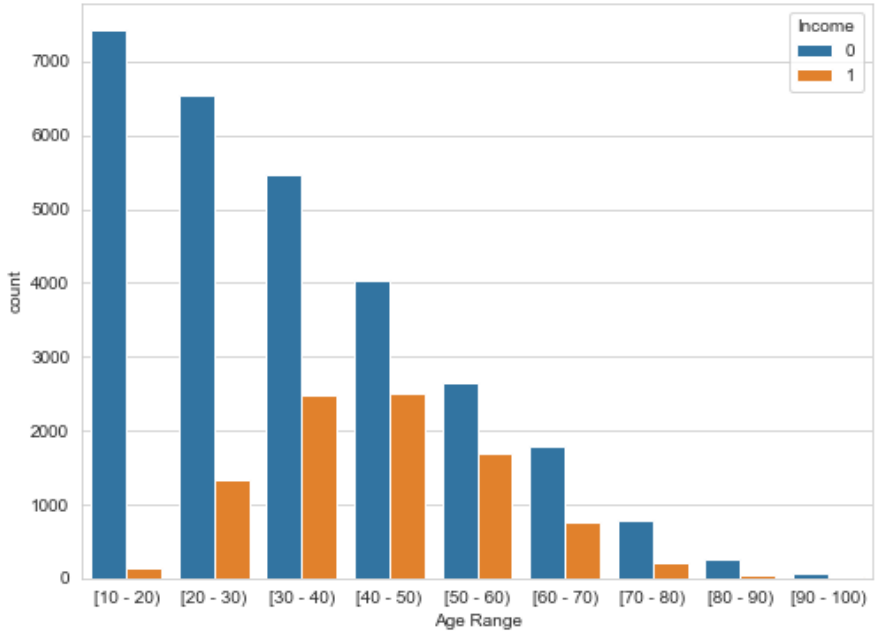
2. The target variable had typo where I got 4 classes instead of two. Basically the labels were correct, just had a (.) as the typo

3. The target variable was also imbalanced with 76% records for <50K and 24% for >50K

4. Variables didn't have much variability which added bias in the final models

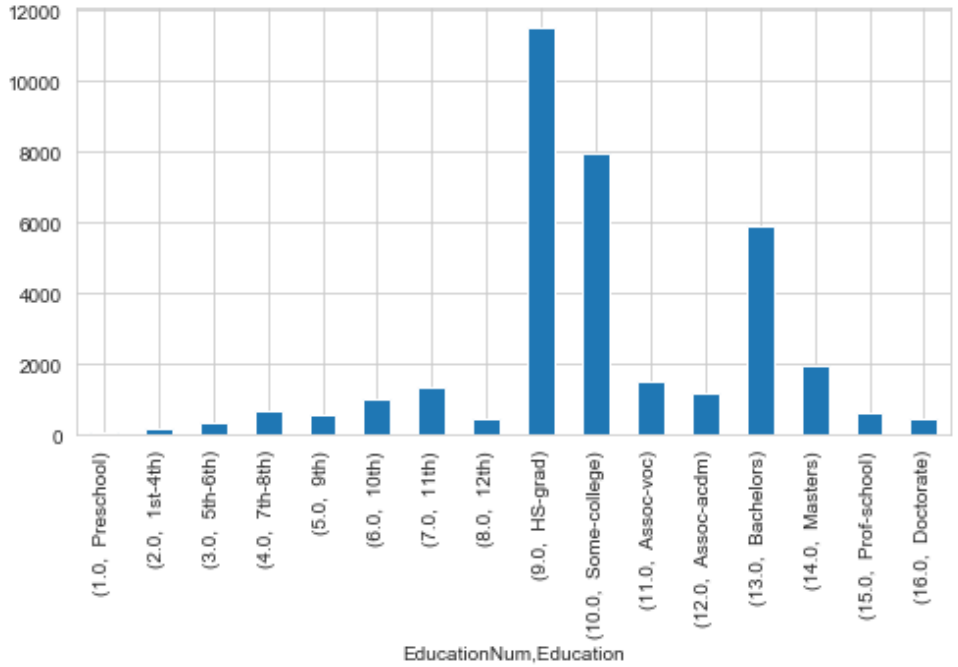
Data Exploration

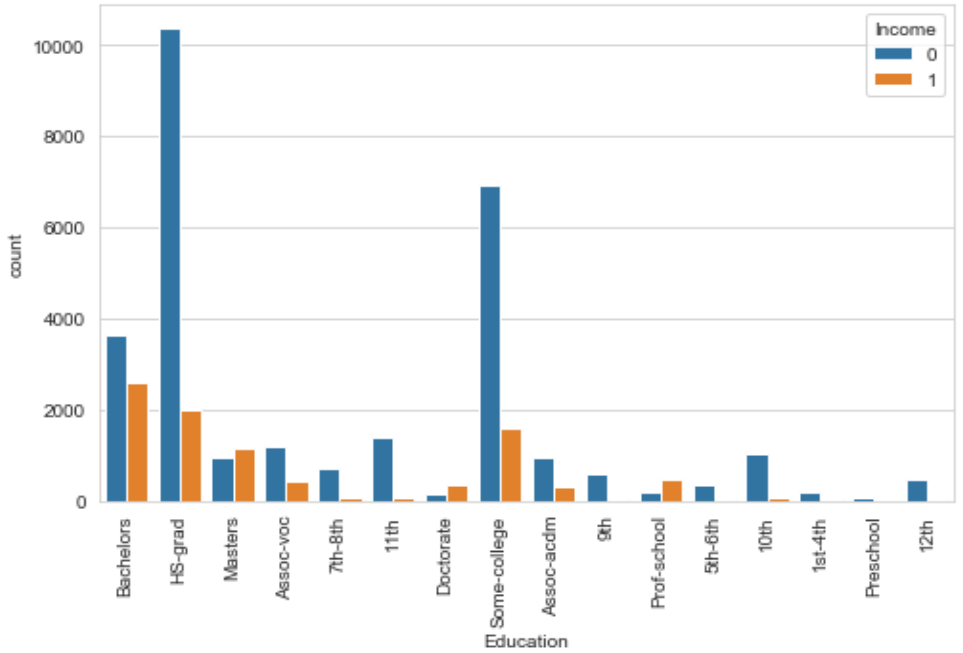




The age feature describes the age of the individual. Majority of ages were between 25 and 50 years. I created age range bins to have a visual representation of how income is affected by age, there is a significant amount of variance between income >50k & <=50k between the age groups.

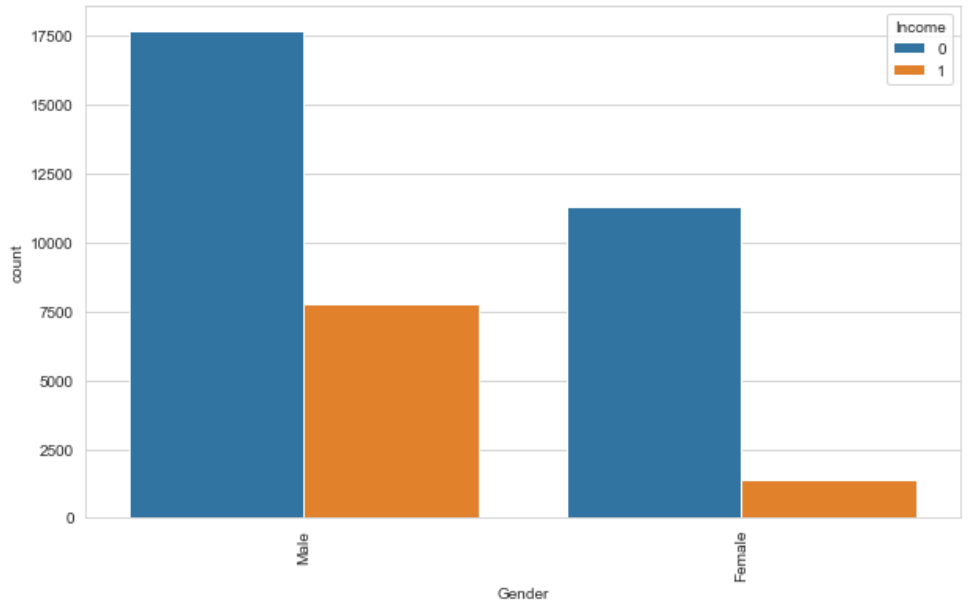
Data Exploration

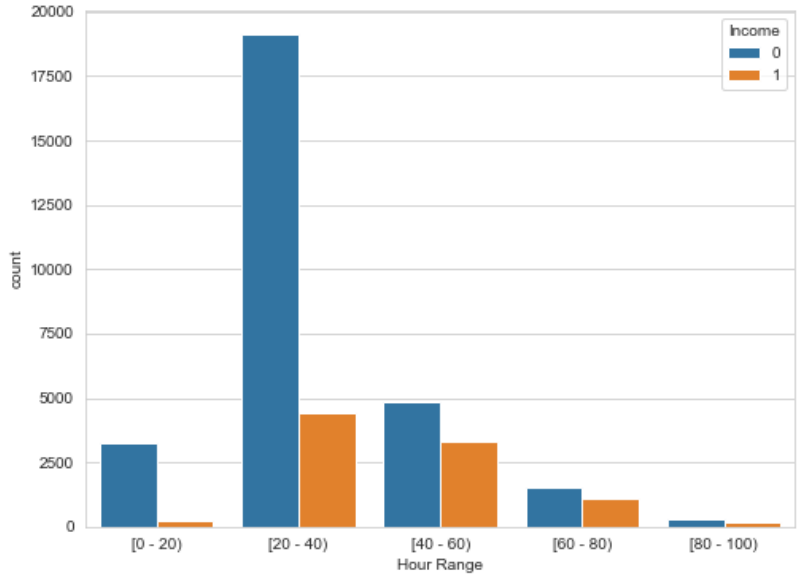




Most of the individuals in the dataset have a high school education while only a small portion have masters and doctorate. Individuals with Higher level of education have greater chances of earning >50k, which is clearly shown by individual's with a Masters or a doctorate degree.

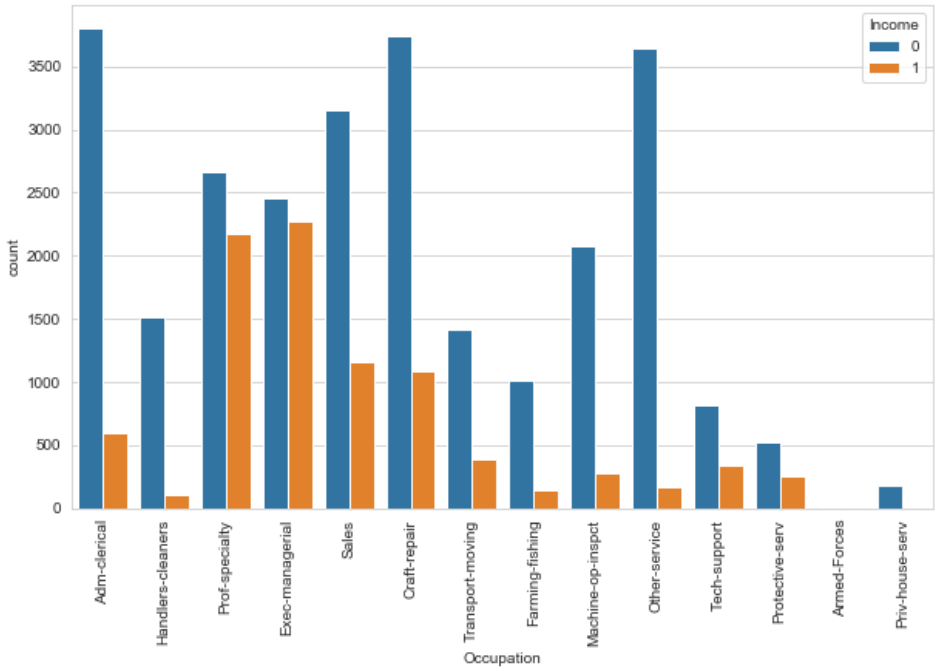
Data Exploration

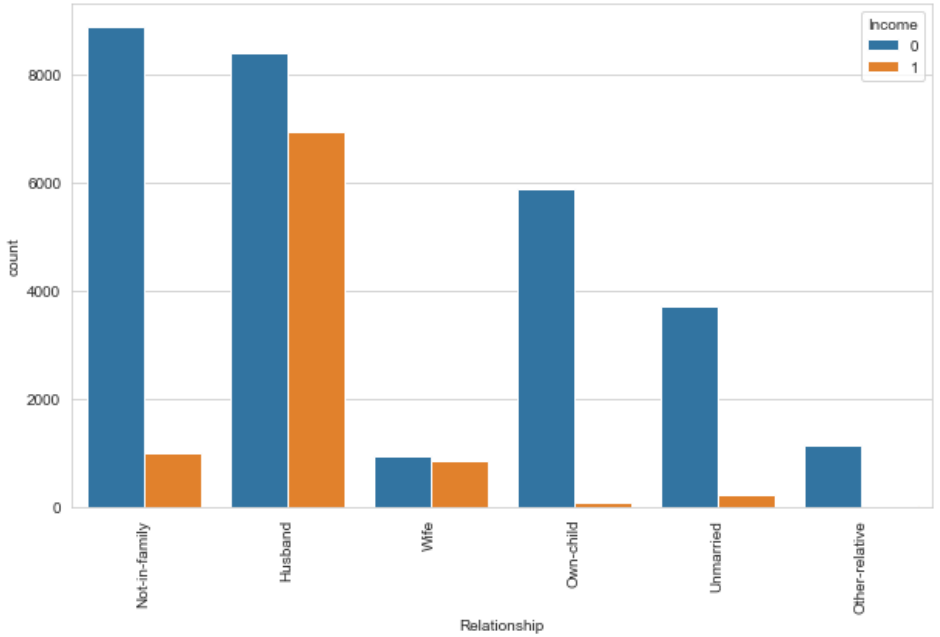




Almost double the sample size of males compared to females for gender variable, Males have a higher chance of earning >50K than females. Most individuals fall into 30-40 hrs/week bracket, which is representative of real world situation and the the higher the age goes percentage difference between earning >50K and <=50K also decreases.

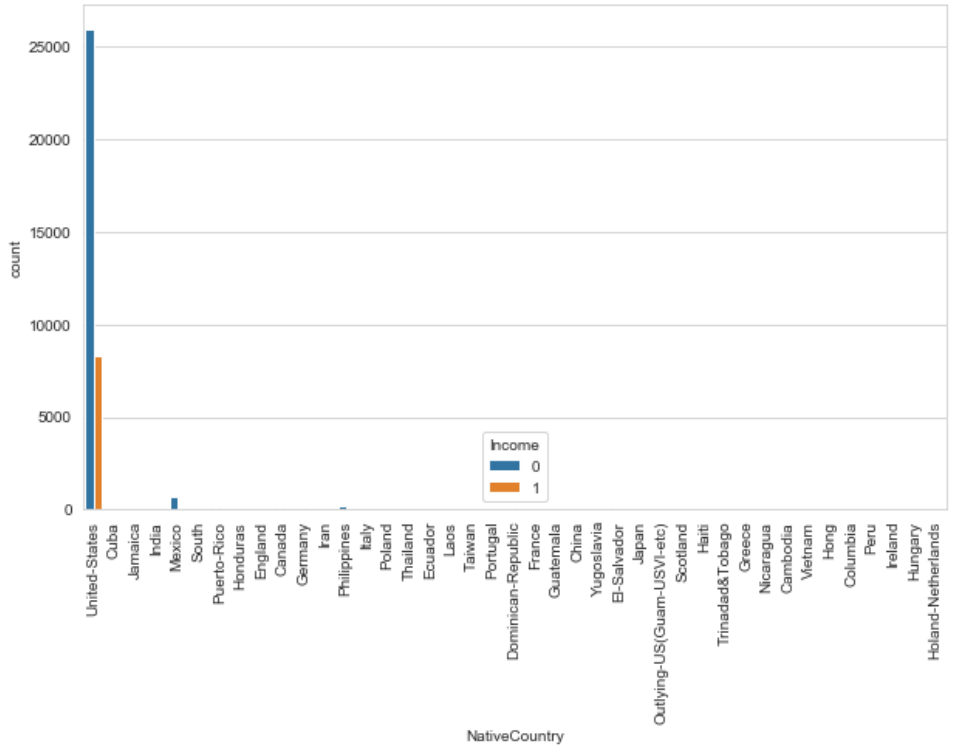
Data Exploration





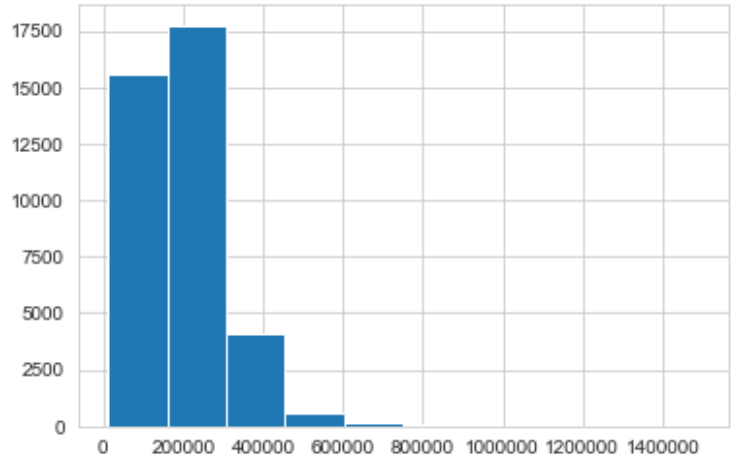
There is good distribution for occupation, one interesting thing to note here is that if an individual Is in a clerical, other-services, and handler - cleaner role they are most likely going to have a hard time earning >50K. For relationship column if an individual is a husband or a wife they are most likely earning >50K.

Data Exploration

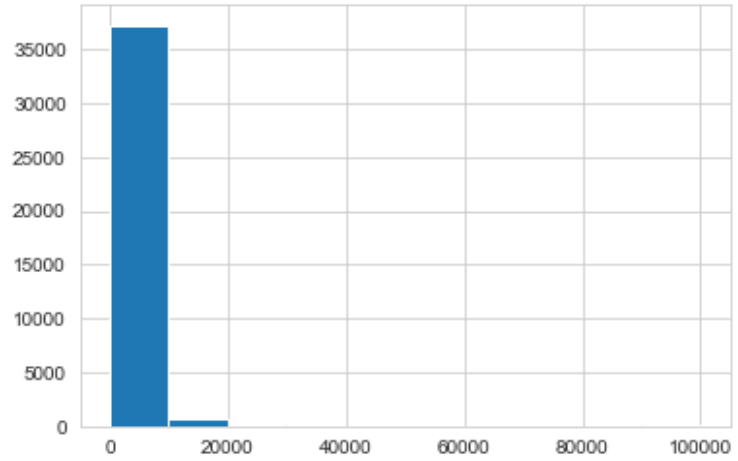


NativeCountry column is heavily skewed with over 75% of individuals representing United-States. For Other countries it looks like no one is earning >50K. Well that is obviously not true, **Assumption:** *we probably lack enough samples for other countries and also there is currency difference. It would be interesting to see if the currencies have been normalized on the same scale or not.*

Data Exploration

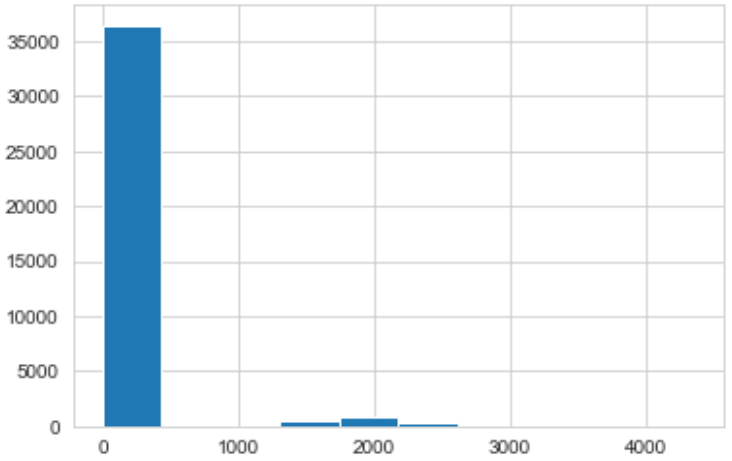


fnlgwt



CapitalGain

These three variables are the least significant ones in the dataset, for example in CapitalGain and CapitalLoss most of the records have 0 , this is not a significant information for the model



CapitalLoss

Feature Engineering

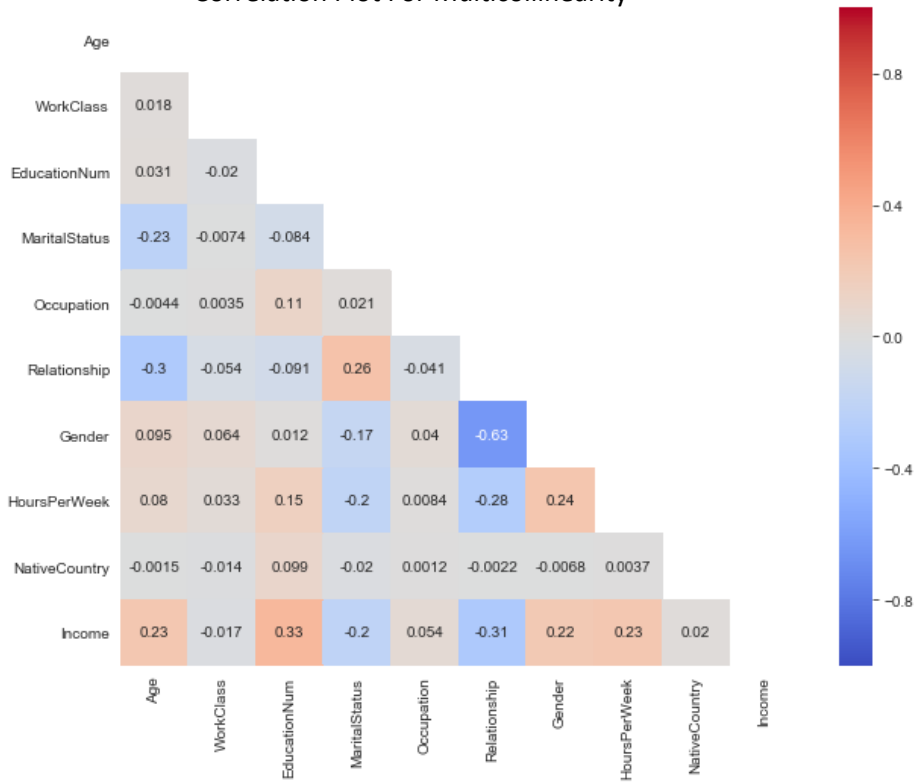
1. *Since there were lots of missing values in the dataset, my first though was to drop the rows with any nulls, That didn't work well since the dropped rows also dropped the class label 1.*

2. *Then I used Iterative Imputer to impute the missing values using similarity from 2-3 nearest neighbors.*

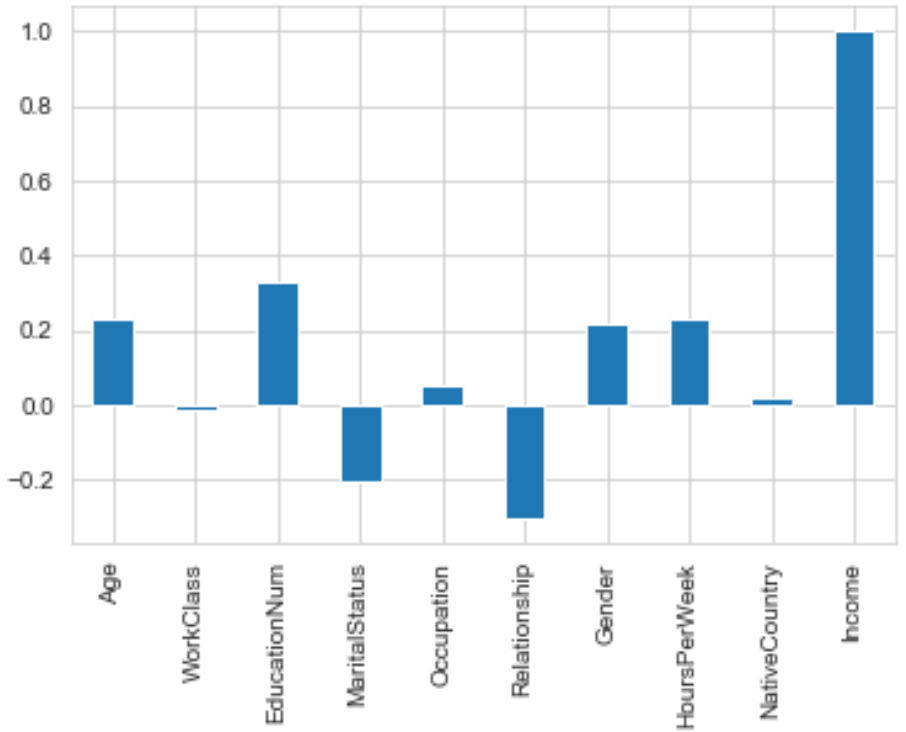
3. *Since most of the variables were categorical, I used LabelEncoder to convert them into numerical attributes.*

Insights

Correlation Plot For Multicollinearity



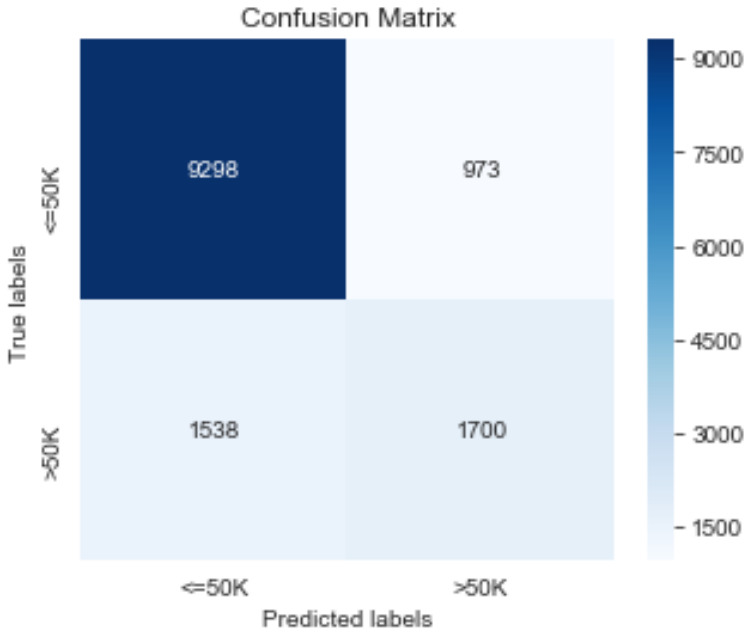
Correlation with Income



I performed a chi-square test to assess the statistical significance among categorical variables, and it stated that each of the variables are important, But on running a correlation plot I was able to see that I should at least remove Relationship because it had a –ve correlation with Gender.

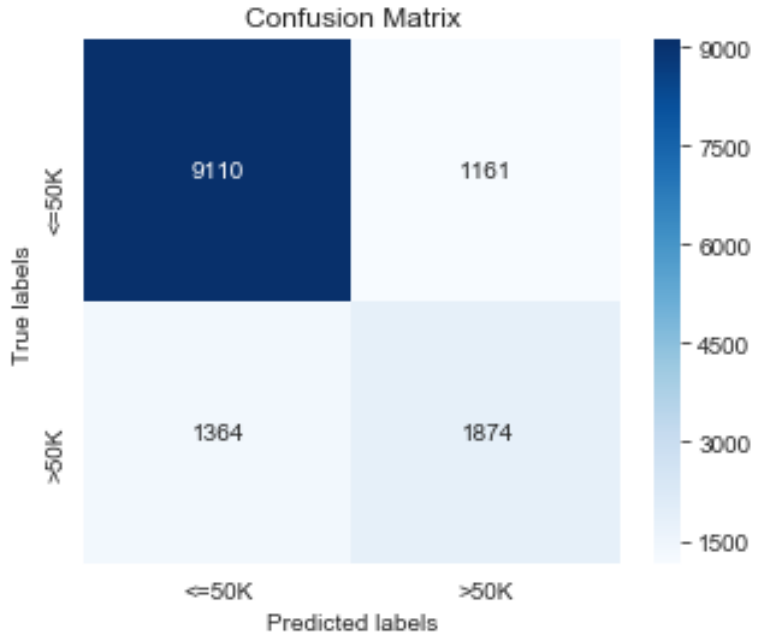
I also plotted a correlation of variables with the target variable, just to see the most important features

Initial Model Performance



I chose to run the RandomForest Classifier model, with all the features except insignificant ones (fnlgwt,CapitalGain,CapitalLoss). My model Overfitted the data with Training Score: 0.96 Testing Score: 0.81, If I would have kept this model it would have been very hard for it to generalize to new examples which would be really bad for the startup.

Final Model Performance



I Iteratively ran 5 other Classifier model (Random Forest, Logistic regression with regularization, XGBoost, Ridge Classifier, Balanced Random Forests), with some features removed, Among these models XGBoost classifier gave me a closer result to what I was expecting from my models, which was basically to not overfit and to reduce misclassification rate of individuals who have earnings '<=50K' into '>50K'.

Model Metrics & Conclusion

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.89	0.88	10271
1	0.62	0.58	0.60	3238
accuracy			0.81	13509
macro avg	0.74	0.73	0.74	13509
weighted avg	0.81	0.81	0.81	13509

Since this was a slightly imbalanced dataset, I didn't use accuracy, precision, and auc as my metric and instead went with highest f1 score for class label 1 (>50K). Since there is a high cost associated with misclassifying an individual when they are earning <=50K and our model classifies them as someone who is earning >50K. Based on that metric XGB Classifier worked best for us. Since it had the lowest misclassification of individuals when they are earning <=50K and our model classifies them as someone who is earning >50K, i.e. 1161.