US Census Data - Its Association with Income and Other Related Issues

Tim Holland, Chuong Tran

Northeastern University

Abstract

This project involves researching and examining different types of data mining methods to use on a US Census dataset provided by Kaggle. More specifically, the goal is to come up with a good model that can help to predict given a certain income status of a person, namely to predict whether that person can earn more or less $50k per year. The proposed methods to apply are Decision Tree, Logistic Regression, Random Forest, and K-nearest neighbors. The final key point of this project will be applying these methods to answer the most important question that needs to be addressed using this dataset: which features are most significant in determining the income earned by a person? At the end, marital status, education number, and capital gain are the biggest three features can help to answer that proposed question given the assistance from Extra Tree Classifier to calculate the relative importance of each feature.

Keywords: Decision Tree, Logistic Regression, Random Forest, K-nearest neighbors

1. Introduction

Kaggle provides a free and open resource for data mining and gives students plenty of options to choose from. US Census data is one of them and it contains many features related to demographic characteristics such as age, race, capital gain, education, etc. Specifically, it includes a collection of 51 state samples. In this current economy, many people continue to struggle to make a living. But even those who are well off remain curious as to what factors play into their financial well-being. Luckily, a lot of data has been collected each year for scientists and researchers to use to make prediction and help improve quality of life. Thus, by using that plethora of data, a lot can be learned about how a person's earned income can be predicted based on current conditions. The question that needs to be addressed here is this: what is the most important factor that contributes to an income earned by a person? More

importantly, we will investigate whether marital status and years of education matter in terms of income earned using various data mining techniques: K-Nearest Neighbors, Decision Tree, Logistic Regression and Random Forest.

2. Problem Definition

2.1 Task Definition

This study will examine the various features that the dataset has to offer and from there the best features that can help to study this problem will be selected. The input will be from the selected 9 features and the output will predict whether a person can make either over 50K per year or at most 50K per year. Thus, the overall goal for this project is to come up with good models that can help to predict an income earned by a person using the selected features.

2.2 Algorithmic Specifications

2.2.1. Decision Tree (CART)

Decision Tree was chosen as one of the classifiers to use because it provides the reasoning behind each split chosen in an efficient manner. It can handle both the continuous and categorical data. Plus, decision tree is a great way to visualize data and help people understand how classes are classified. It works well with simple binary classification problems like the census data's class label of income being over fifty thousand or under fifty thousand per year. However, some additional preprocessing is required to use Decision Tree with categorical features like with some of the census features such as marital status and pruning is needed to minimize overfitting.

2.2.2. Logistic Regression

Logistic Regression provides a probably of a classification being true or not and, like Decision Tree, works well with binary classification problems. This is important for predictions because it provides a quick measure of confidence in the assigned class. It's also an efficient classifier with low variance and is not prone to overfitting. Since the outcome is discrete with two classes specifically, it's another reason why Logistic Regression should be a perfect fit for this project.

2.2.3. Random Forest

Random Forest provides a good benchmark among the classifiers chosen because of its accuracy and ease of use over other classifiers. Due to the work put in to Scikit-Learn's Random Forest implementation it's also relatively efficient, so it has a minimal cost of implementation. It is expected that Random Forest to provide the best results, but it is desirable to retain some of the explanatory results of Decision Tree and Logistic Regression.

2.2.4. k-NN

k-NN is a straightforward way of performing a sanity check on our classification approach. It is a supervised learning method and can be good to use for classification tasks. Plus, It requires choosing the right k-value, and is less powerful and efficient than our other classifiers, but in some cases, k-NN may be all that's required to get a good result. Due to this, it's expected to have more modest accuracy in comparison to the other classifiers.

3. Experimental Methods and Results

3.1 Experimental Methodology

3.1.1 Data Introduction

The data obtained from Kaggle was extracted from the 1994 US Census. There are a total of 32561 samples and 15 features. The features come in with mix types such as continuous and discrete. A number of these features are directly or indirectly relevant to income (see figure 1).

US Census Features		
Feature	*Type*	*Additional information*
age	continuous	
workclass	categorical	
fnwgt	continuous	Irrelevant calculation
education	categorical	
education.num	discrete/continuous	[1-16] (approx. years)
marital.status	categorical	
occupation	categorical	Example: Tech-support
relationship	categorical	Place in family (ex: husband)
race	categorical	Example: Black
sex	categorical	Male or female
capital.gain	continuous	

capital.loss	continuous	
hours.per.week	continuous	Hours worked per week
native.country	categorical	Example: Greece
income	categorical	>50k or <=50k

Figure 1. Features present in the dataset and their descriptions

3.1.2 Data Processing

In order to apply different data mining techniques, the dataset needs to be preprocessed first since missing or malformed data can hurt the results of the classifiers later on. After thoroughly going through the dataset, there are about 2000 missing data values. Fortunately they are not that important and can be eliminated completely since they belong to features that we are not interested in. The missing data is also across a couple of features rather than just one, and as a result it will not be beneficial to use those data.

After removing data, we need to find a way to balance as more data belongs to the less than or equal to 50K class than to the greater than 50K class. To avoid the problem of either overfitting or underfitting, SMOTE has been applied to perform oversampling to create an equal number of instances for the two outcome classes.

After balancing the data, the next step will be changing the data types so that all the input features can have the same type and each proposed method can be applied to them. Each categorical feature is assigned to integer values. For example with the sex feature, male can be assigned to number zero and female can be assigned to one.

The features have to be selected carefully as there are 15 of them and some of them will be irrelevant to this project, such as final weight and occupation, as their input data are either missing or offer no insight to the questions posed. The last step for data processing is applying Extra Trees to select the best features from the 9 selected features that we choose originally: native country, capital loss, capital gain, relationship, sex, race, marital status, education and workclass. The reasoning for picking these features instead of using all 15 is based on outside research that indicated these features should be examined. In addition, the point of including native country is to check whether it contributes or plays any factor in the questions posed. Other features might play an important role as well, but after experimenting with various

combinations of features, the most important features were selected (see figure 2).

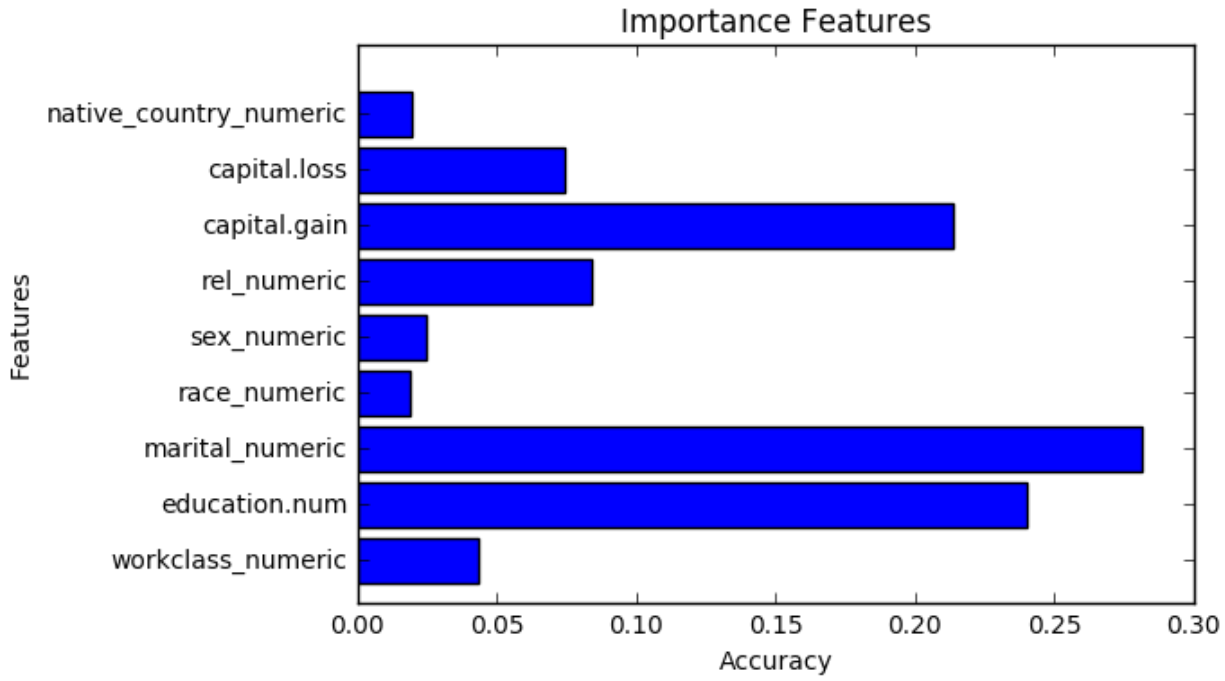

Figure 2. Important features as determined by Extra Trees

3.2 Results and Analysis

The analysis was run both with and without SMOTE to compare the results. To determine the number of estimators parameter for Random Forest, a number of values starting with 1 were tested until it was determined there was no longer any gained benefit, which was 25 (see figure 3). Likewise, for k-NN, values of k were iterated over starting at 1 until no improvement in accuracy was seen, which was 7. Both 'distance' and 'uniform' were tried with k-NN, with no major difference between the two but a very slight advantage with distance, so distance was used. For Decision Tree, since the most important features were already determined using Extra Trees, the number of features was set to 3. The depth was set to 3 to keep the results understandable. A slightly higher accuracy was seen with a higher depth, but the resulting splits didn't add much value.

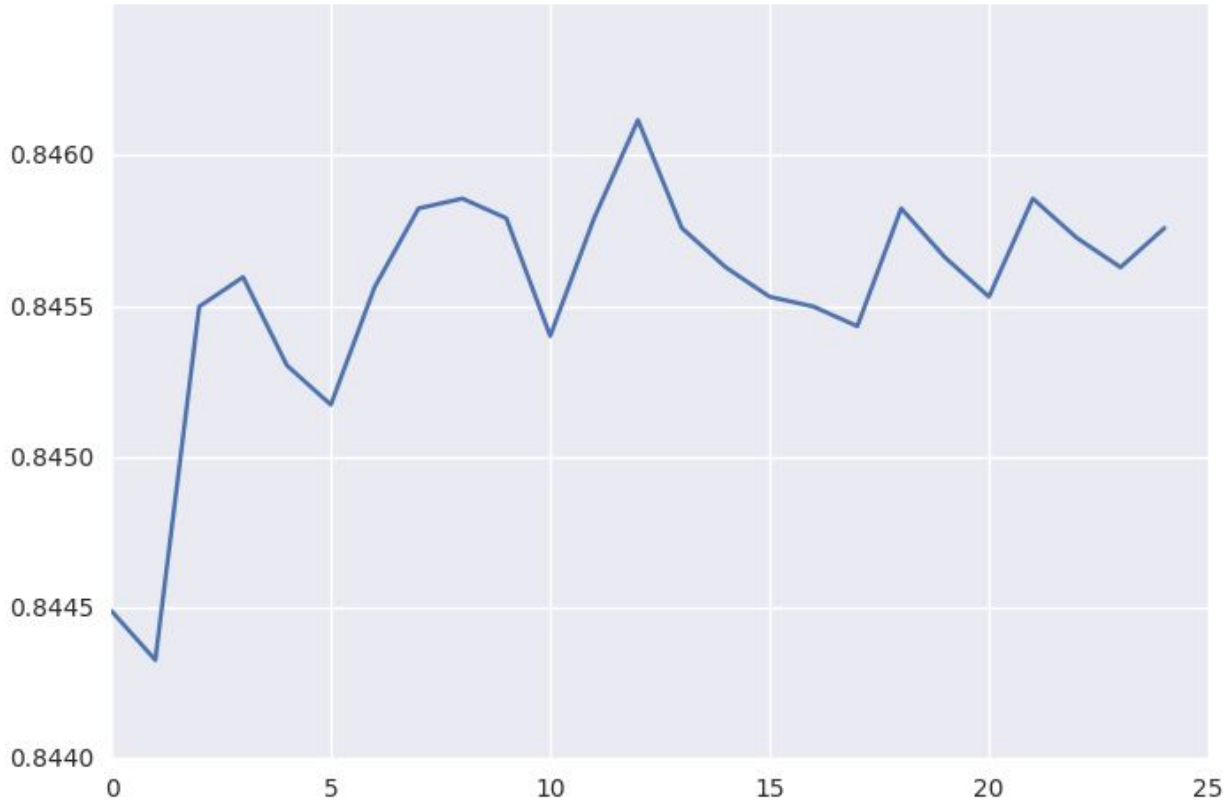

Figure 3. Accuracy of Random Forest using N-estimators.

K-Fold Stratified Cross-Validation with a k equal to 5 was used to ensure that the training and test sets were appropriate. This value was chosen because it produces a good balance between the size of the training set and the size of the test set. Accuracy was higher without SMOTE and the majority class of income <=50k had a better F1-score, but SMOTE slightly improved the F1-score for income >50k. Overall it appears SMOTE did little to improve the results since correctly classifying income >50k is no more important than correctly classifying <=50k in this analysis. An example of the difference between using SMOTE and not can be seen in the confusion matrices for Random Forest in figure 4. Random Forest edged ahead of the other classifiers in most metrics, especially in its F1-scores, but not by a large margin. k-NN had unusually high accuracy when using SMOTE, but its variance was 0.06 in comparison to a variance of 0.01 for other methods, indicating a certain degree of unreliability. For a side-by-side comparison of the metrics between the classifiers, see figures 5 and 6.

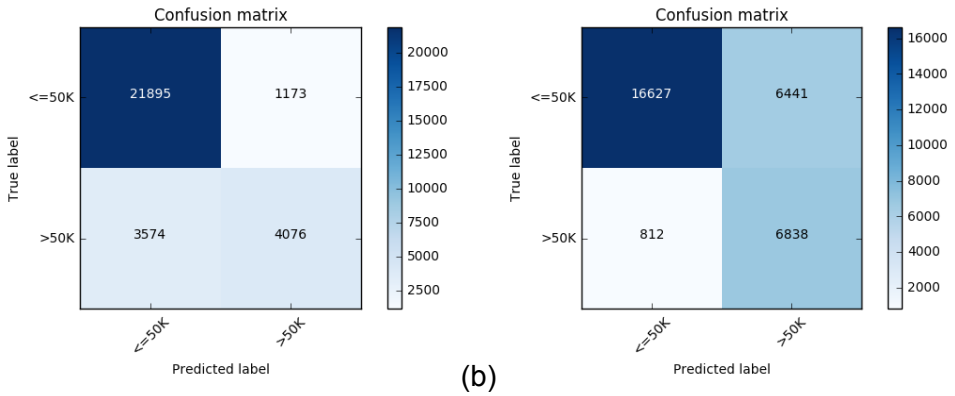

Figure 4. (a) Confusion Matrix for Random Forest without using SMOTE. (b) Confusion Matrix for Random Forest using SMOTE

Metrics with SMOTE	Decision Tree	Random Forest	Logistic Regression	KNN
Accuracy	0.76	0.76	0.74	0.81
F1-Score	0.77	0.78	0.76	0.81
F1-Score (>50k)	0.63	0.65	0.63	0.62
F1-Score (<=50k)	0.82	0.82	0.80	0.88

Figure 5. Comparing each model using one common dataset using SMOTE

Metrics without SMOTE	Decision Tree	Random Forest	Logistic Regression	KNN
Accuracy	0.84	0.85	0.83	0.82
F1-Score	0.83	0.84	0.82	0.82
F1-Score (>50k)	0.62	0.63	0.62	0.63
F1-Score (<=50k)	0.90	0.90	0.89	0.88

Figure 6. Comparing each model using one common dataset without SMOTE

Decision Tree determined the marital status was the most important feature to split on. For people who are for whatever reason currently unmarried (i.e. have a marital numeric value under 5), the next important feature for classifying income was determined to be capital gain. For people who are currently married, the next important feature was determined to be

education, with having some amount of college education indicating an income over 50k. For those without a college education, having more capital gain was the deciding factor. The correlation between the three features examined and the two income groups can be seen at a glance in figure 8.

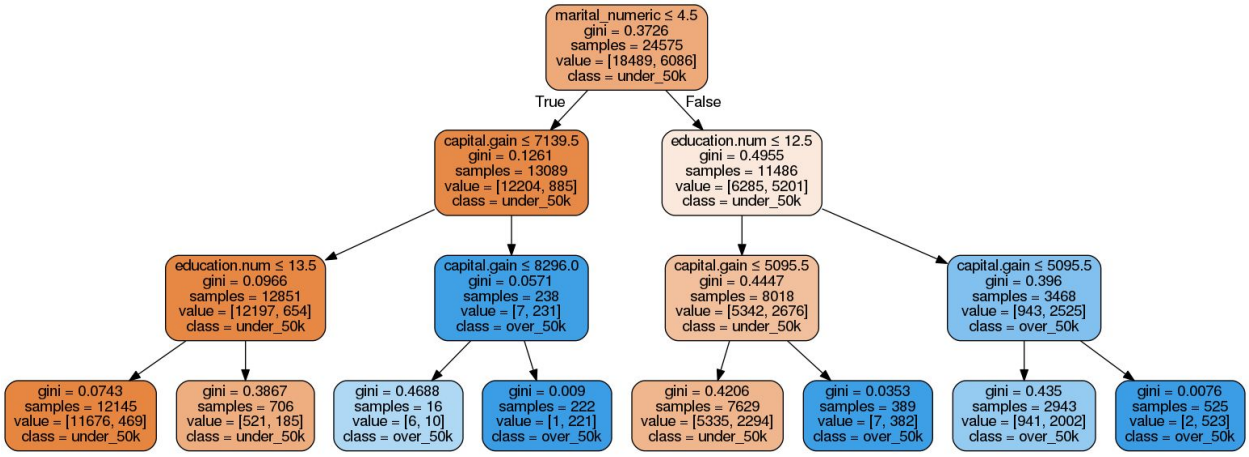

Figure 7. Decision Tree at a depth of 3.

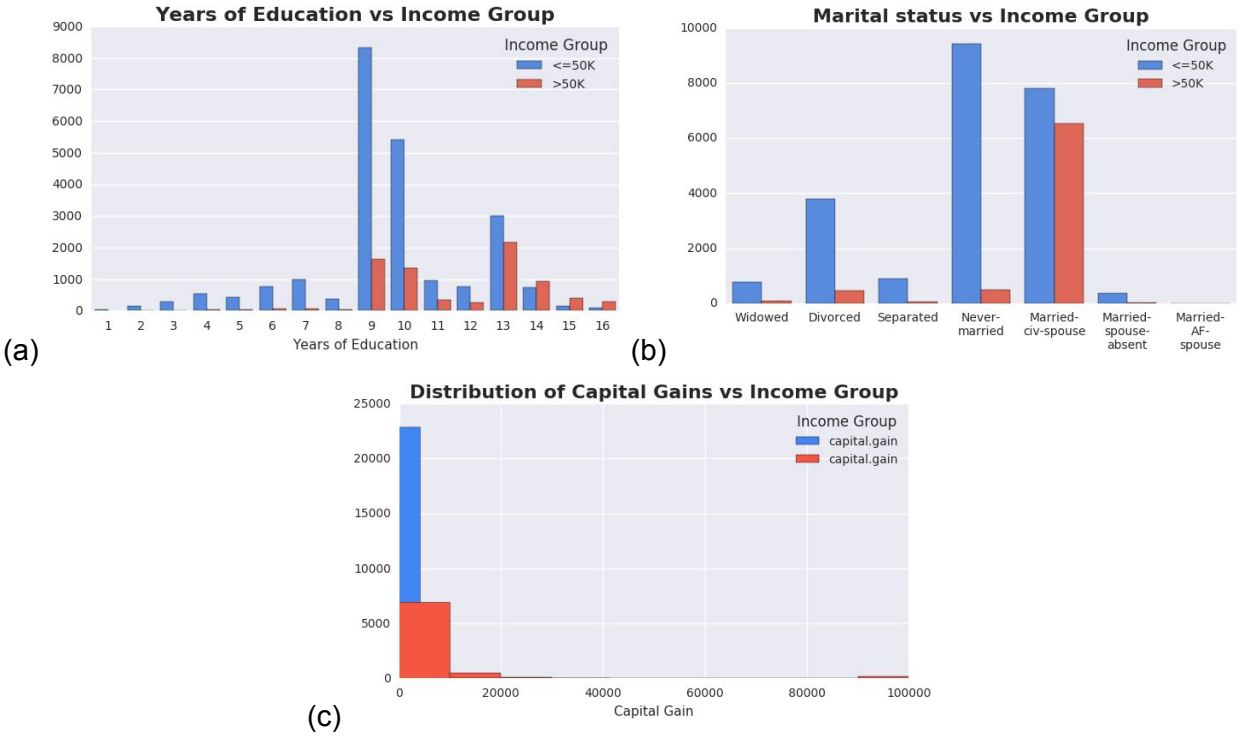

Figure 8. (a) Approximate of education by income. (b) Marital status by income. (c) Capital gains by income

3.3 Discussion

Surprisingly, race and country of origin did not prove to be as significant as initially anticipated. When compared to other features during the Extra Trees analysis these two features displayed little importance. Part of this may be due to the nature of census data and collecting population statistics in general in the United States - undocumented immigrants may be very hesitant to speak to people representing the US government, which acts as a significant factor in data collection. Due to this limitation and the nature of visa programs in the US, it's possible that race and country of origin become diminished features in census data. Education and capital gain are known to be correlated with income, so having these features determined as important is expected. To some degree it's known that marital status impacts income, although its relative importance in this analysis was unexpected.

Without using SMOTE, the methods all showed an improvement on classifying income under 50k, and with SMOTE showed an improvement on classifying income over 50k. k-NN showed a higher degree of variance, although it retained accuracy when applying SMOTE. k-NN also may have issues with memory use and computation speed if a larger dataset were to be applied, so for this case it works but limits its future applications. Overall, without using SMOTE, Random Forest edged out the other methods and showed a good degree of balance of precision and recall, as shown by its F1-score. Random Forest takes a more extensive approach to exploring subsets of data with a form of bagging, which explains why it outperforms Decision Tree. Logistic Regression and Decision Tree both showed good results, as well, meaning their use in explaining predictions can still be useful. Their good performance makes sense given their strength in binary classification problems. Variance, accuracy, F1-scores, and confusion matrices all provided a good overall picture of the performance of each method.

One weakness in the features used on the methods is that marital status can't be cleanly converted to a continuous feature, but needed to be done since categorical features aren't handled well by the tree methods and Logistic Regression. Given the relative importance of marital status in this analysis it's possible that this limited the strength of the results to some degree.

4. Future Work

It is possible to utilize richer data sources and take different approaches with the dataset. Additional feature selection methods should be used to couple with Extra Trees to ensure that

the important features can be selected correctly. For now, this project relies heavily on only one classifier to select the best features possible. Depending on how data is processed, it can leave a big impact on the steps that follow it in the project. Recursive Feature Elimination is good alternative approach that can be coupled with Extra Trees. Thus, it can be used in the future in a side-by-side comparison. Another possibility is that different classifier methods can be taken with this dataset. Some features are excluded in this analysis due to the focus being on certain features. For example, age is not included since it has relatively low impact on classification results when including other features. It should be investigated in further detail. Furthermore, other ensemble methods can also be applied to improve the results by using either stacking or bagging.

5. Conclusions

Based on the results obtained from each methods, it is safe to conclude that the three selected features (marital status, education number and capital gain) from Extra Tree Classifier are indeed somewhat important. The accuracy is very close between the four proposed methods: K-nearest neighbors, Decision Tree, Logistic Regression and Random Forest. More importantly, using SMOTE does not guarantee to improve the accuracy for those methods as they have been tested with it and without it and the accuracy does not change that drastically. These methods were evaluated using the generated classification report from Scikit Learn, confusion matrix, and K-fold cross validation to ensure that they can be used to support the proposed questions. These are just proposed methods and means of evaluation for this particular dataset and there are definitely many additional ways to improve and help make this project better in term of feature selection approach, hyperparameter selection, and the evaluation for each method.

6. References

Ronny Kohavi and Barry Becker, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 1994.

Ronny Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.