Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques

Chet Lemon (A10895241)
Chris Zelazo (A10863450)
Kesav Mulakaluri (A10616114)

Abstract -- For this assignment, we examine the Census Income dataset available at the UC Irvine Machine Learning Repository. We aim to predict whether an individual's income will be greater than $50,000 per year based on several attributes from the census data.

Introduction

The US Adult Census dataset is a repository of 48,842 entries extracted from the 1994 US Census database.

In our first section, we explore the data at face value in order to understand the trends and representations of certain demographics in the corpus. We then use this information in section two to form models to predict whether an individual made more or less than $50,000 in 1994. In the third section, we look into a couple papers written on the dataset to find out what methods they are using to gain insight on the same data. Finally, in the fourth section, we compare our models as well as that of others in order to find out what features are of significance, what methods are most effective, and gain an understanding of some of the intuition behind the numbers.

Exploratory Analysis

The Dataset

The Census Income dataset has 48,842 entries. Each entry contains the following information about an individual:

- **age**: the age of an individual
  - Integer greater than 0
- **workclass**: a general term to represent the employment status of an individual
  - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt**: final weight. In other words, this is the number of people the census believes the entry represents..
  - Integer greater than 0
- **education**: the highest level of education achieved by an individual.
  - 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**: the highest level of education achieved in numerical form.
  - Integer greater than 0
- **marital-status**: marital status of an individual. Married-civ-spouse corresponds to a civilian spouse while Married-AF-spouse is a spouse in the Armed Forces.

    - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- **occupation**: the general type of occupation of an individual
    - 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**: represents what this individual is relative to others. For example an individual could be a Husband. Each entry only has one relationship attribute and is somewhat redundant with marital status. We might not make use of this attribute at all
    - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race**: Descriptions of an individual's race
    - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **sex**: the biological sex of the individual
    - Male, Female
- **capital-gain**: capital gains for an individual
    - Integer greater than or equal to 0
- **capital-loss**: capital loss for an individual
    - Integer greater than or equal to 0
- **hours-per-week**: the hours an individual has reported to work per week
    - continuous.
- **native-country**: country of origin for an individual
    - 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.
- **the label**: whether or not an individual makes more than $50,000 annually.
    - <=50k, >50k

The original dataset contains a distribution of 23.93% entries labeled with >50k and 76.07% entries labeled with <=50k. We split the dataset into training and test sets while maintaining the above distribution. The following graphs and statistics pertain to the training set.

Label	Number	Percentage
<= 50k	7841	24.1
> 50k	24720	75.9

To gain insights about which features would be most helpful for this assignment, we look at the feature and the distribution of entries that are labeled > 50k and <= 50k. We do this in hopes to identify features that provide little information in order to simplify our model's complexity and runtime.

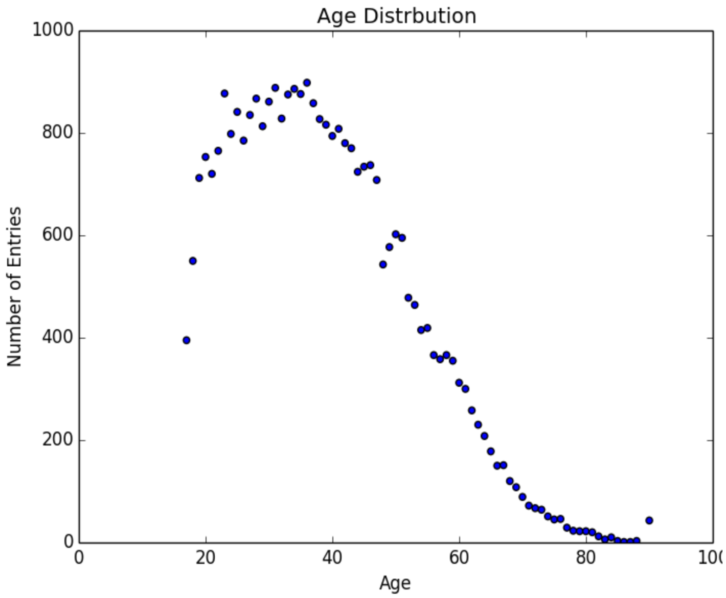


Fig. 1 Age Distribution

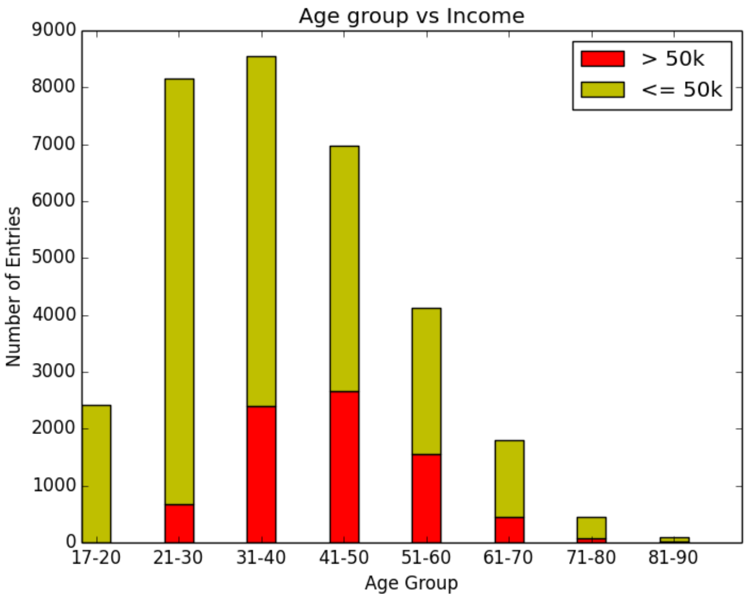
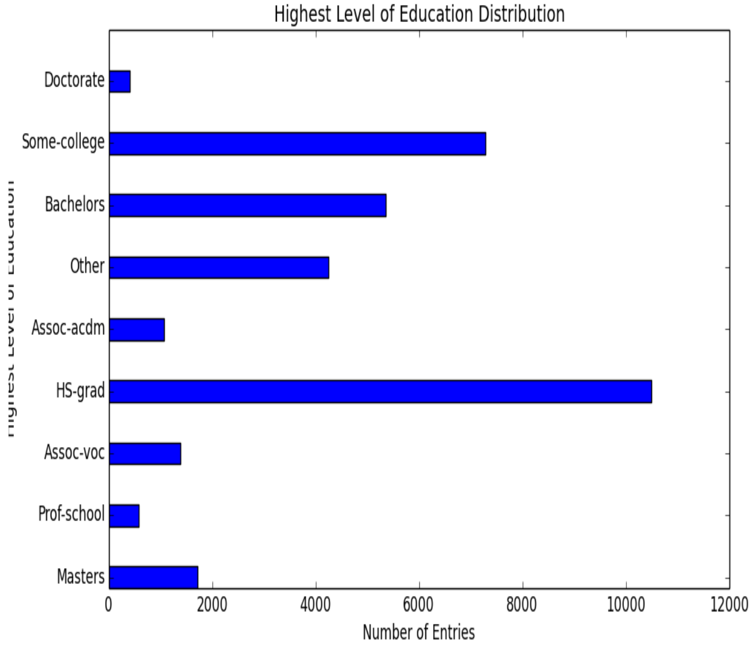


Fig. 2 Age Group vs Income

The age feature describes the age of the individual. **Figure 1** shows the age distribution among the entries in our dataset. The ages range from 17 to 90 years old with the majority of entries between the ages of 25 and 50 years. Because there are so many ages being represented, we bucket the entries into age groups with intervals of ten years to present the data more concisely as seen in **Figure 2**. Looking at the graph, we can see that there is a significant amount of variance between the ratio of >50k to <=50k between the age groups. The most interesting ratios to note are those of groups 17-20, 71-80, and 81-90 where there is almost no chance to have an income of greater than $50,000. The ratio of entries labeled >50k to <=50k for age groups 21-30, 31-40, 41-50, and 51-60 vary significantly as well.



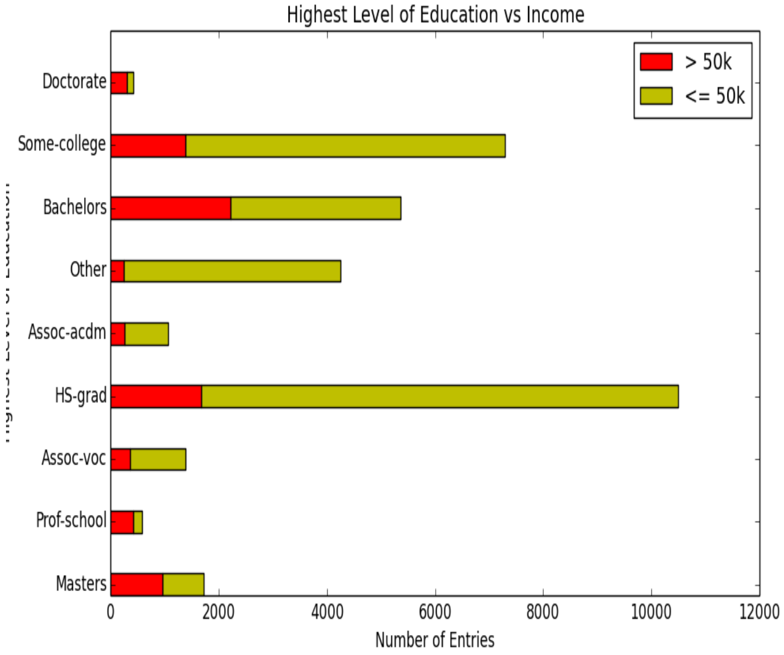
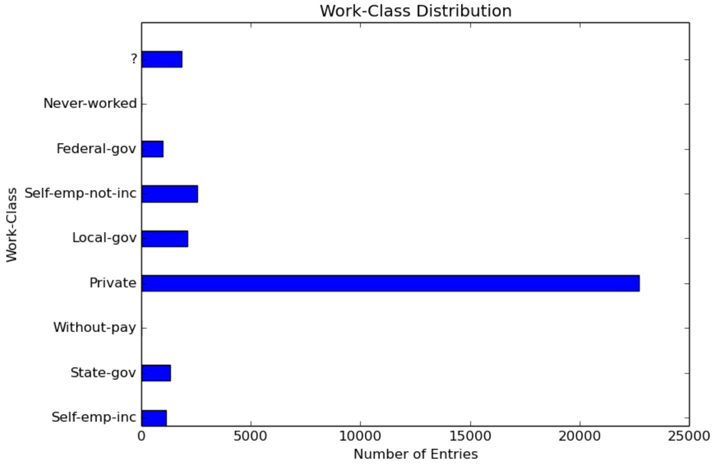


Fig. 3 Highest Level Education Distribution

Fig. 4 Highest Level Education vs Income

The education feature describes the highest level of education of each individual in the dataset. **Figure 3** shows the distribution of the different levels of education among individuals in the dataset. The Other group represents Preschool through 12th grade. Most of the individuals in the dataset have at most a high school education while only a small portion have a doctorate. We think this is a fair representation. **Figure 4** shows the relationship between the highest level of education and the number of people labeled >50k and <=50k. For the most part, a higher level of education is correlated to a higher percentage of individuals with the label >50k. One interesting statistic to note is the ratio of individuals labeled >50k to <=50k is almost the same between those that have a doctorate and those that went to a professional school (Prof-school).



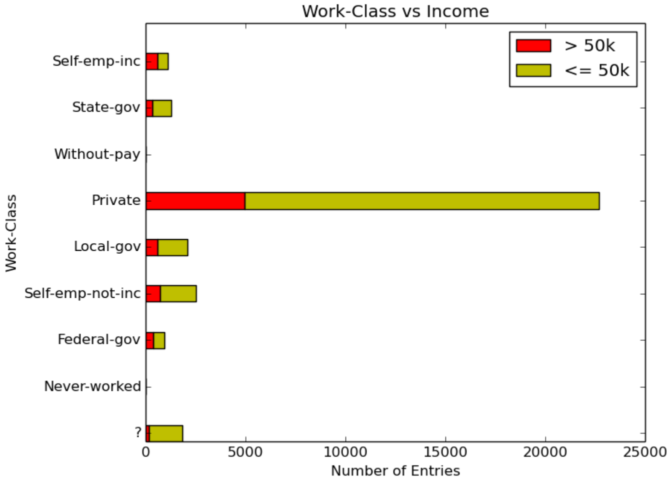
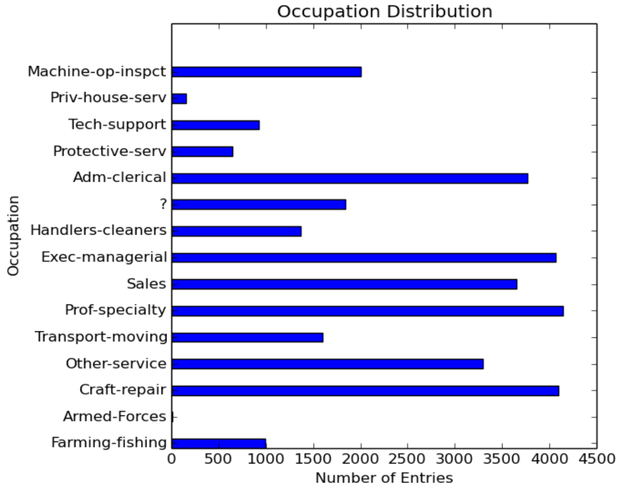


Fig 5. Work Class Distribution

Fig 6. Work Class vs Income

As seen in **Figure 5**, the majority of the individuals work in the private sector. The one concerning statistic is the number of individuals with an unknown work class. The probabilities of making above $50,000 are similar among the work classes except for self-emp-inc and federal government. Federal government is seen as the most elite in the public sector, which most likely explains the higher chance of earning more than $50,000. Self-employed-incorporated implies that the individual owns their own company, which is a category with an almost infinite ceiling when it comes to earnings. The complete Work Class vs Income graph can be seen in **Figure 6**.



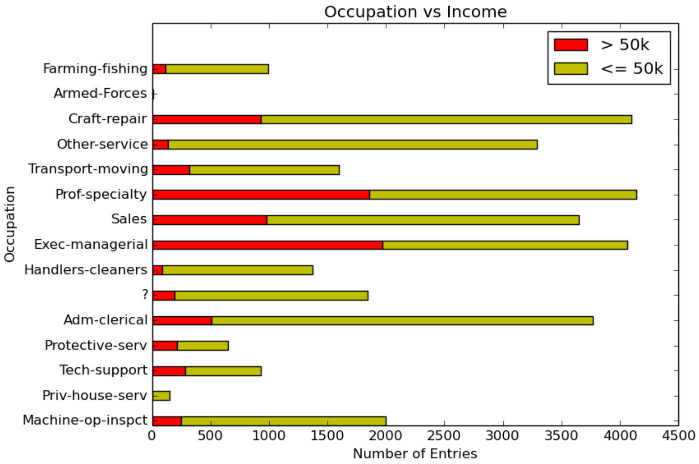
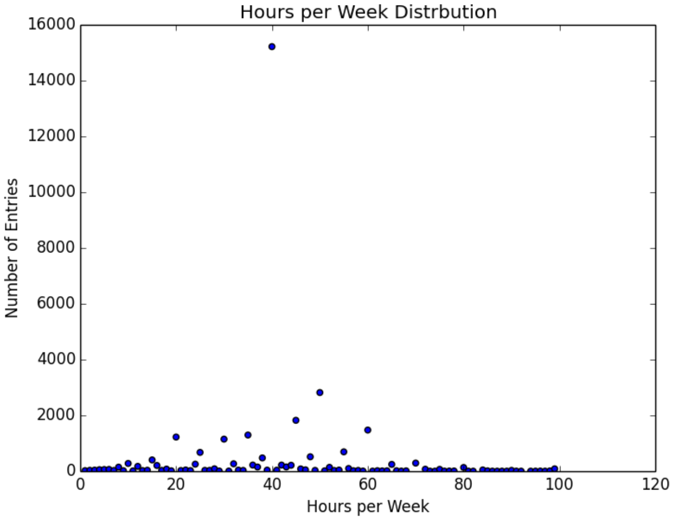


Fig. 7 Occupation Distribution

Fig. 8 Occupation vs Income

As seen in **Figure 7**, there is a somewhat uniform distribution of occupations in the dataset, disregarding the absence of Armed Forces. However, looking at **Figure 8 Occupation vs Income**, exec-managerial and prof-specialty stand out as having very high percentages of individuals making over $50,000. In addition, the percentages for Farming-fishing, Other-service and Handlers-cleaners are significantly lower than the rest of the distribution. The one concerning statistic looking at **Figure 8** is the high number of individuals with unknown occupations.



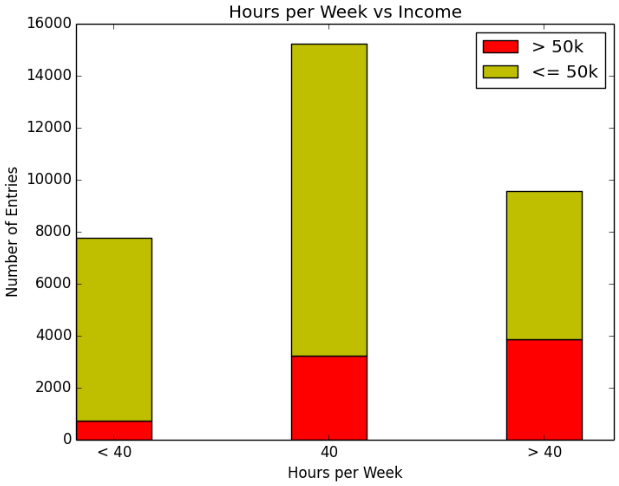


Fig. 9 Hours/Week Distribtution

Fig. 10 Hours/Week vs Income

Looking at the distribution in **Figure 9**, the vast majority of individuals are working 40 hour weeks which is expected as the societal norm. Regardless of the nonuniform distribution, **Figure 10** shows that the percentage of individuals making over $50,000 drastically decreases when less than 40 hours per week, and increases significantly when greater than 40 hours per week.

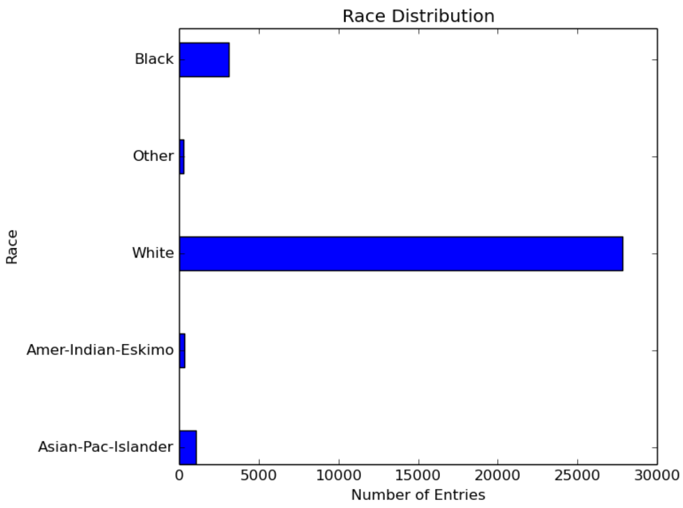


Fig. 11 Race Distribution

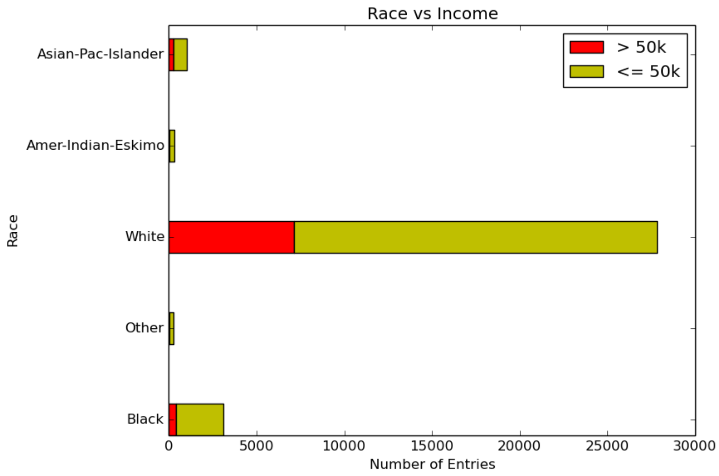


Fig. 12 Race vs Income

Looking at the **Figure 12**, it seems like the feature could be useful in our prediction model, as Whites and Asians have a larger percentage of entries greater than $50,000 than the rest of the races. However, the sample size of Whites in the dataset is disproportionately large in comparison to all other races. The second most represented group is Blacks with less than 5000 entries. The lack of equal distribution caused us to consider not utilizing this attribute in our prediction model.

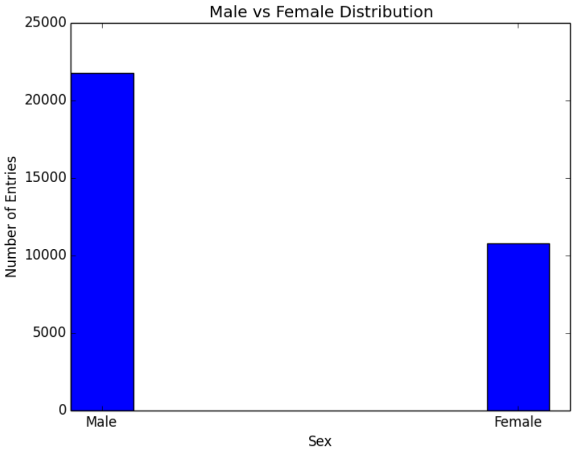


Fig. 13 Male vs Female Distribution

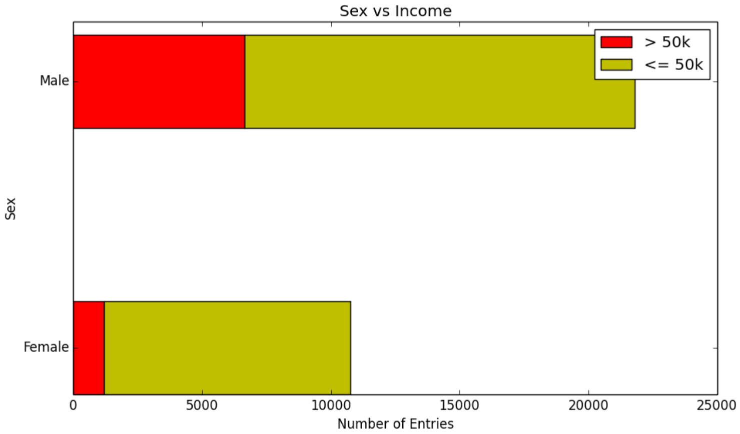


Fig. 14 Sex vs Income

In **Figure 13**, we can see that there is almost double the sample size of males in comparison to females in the dataset. While this may not affect our predictions too much, the distribution of income can. As seen in **Figure 14**, the percentage of males who make greater than $50,000 is much greater than the percentage of females that make the same amount. This will certainly be a significant factor, and should be a feature considered in our prediction model.

Removal of Samples
When we began testing some of our models, we had issues creating default values for many of the samples with unknown or missing values which caused errors in calculation and skewed results. We ended up removing any data sample that had a missing value to improve our results.

Removal of Features
We also opted to not use the features: ‘fnlwgt’, ‘relationships’, and ‘capitalGains/Loss’. These features either were not useful for our analysis or had too much bad data i.e. zero-values, unknown/private values.

Predictive Task
Baselining the Data
As mentioned before, we saw a distribution of roughly twenty-four percent entries labeled with >50k and seventy-six percent labeled with <=50k. In order to establish baseline data for our classifiers, we predicted the majority label *<=50k* for each item. This model gave us the baselines described in **Figure 15** below.

	Training	**Testing**
Error	24.008%	23.623%

Fig. 15 Baseline Data

Predicting of Salaries Greater than $50k/year

Naïve Bayes
The model we started with in order to predict an individual’s salary range was based on Naïve Bayes. In the most simple form, we tried to choose features we felt were independent of one another and not too skewed based on their distribution.

> For example*, the number of samples from individuals identifying as ‘White’ was severely disproportionate to even the sum of all of the rest of the samples in the data set, thus we chose to ignore this feature.*

We combined the features *‘age’*, *‘sex’*, *‘education’*, *‘hoursPerWeek’*, and *‘occupation’* in order to achieve the lowest percent prediction error when evaluating against the test data set.

Naïve Bayes	**Training**	**Testing**
Error	19.893%	20.432%

The above table shows the classification when trained on the unadulterated data points. We also tested grouping data items together in order to make simpler and more conclusive classification. For example, the *'education'* feature had 16 different values, many of which overlapped so we combined values of 'Preschool' to '12th' as they are all consistently low performing. For the same reasoning, we also made ranges of ages and grouped countries together into regions. We speculated that for many of the smaller feature values that seem to follow the same trends, combining them would produce a stronger average leading to more accurate classification the test set was much smaller and had some previously unseen keys. As promising as this seemed initially, grouping proved to only drive our error rates higher and waste our time. The table below reflects this.

Naïve Bayes (Grouped)	**Training**	**Testing**
Error	22.353%	24.128%

Logistic Regression

After our success with the Naïve Bayes classifiers, we decided to try predicting with a Logistic Regression classifier. No combination of features proved successful - even the best error rate was significantly less successful than our simple Bayesian model. The table below represents the features *'age'*, *'education'*, *'maritalStatus'*, *'sex'*, and *'hoursPerWeek'* in the X feature vector and the binary representation of < or >= $50k in the Y vector.

We reason that because our data set is not easily linearly separable, this classifier fails on predicting with these features. If we instead had actual salaries of individuals rather than merely a binary indicator, this model could be successful.

Logistic Regression	**Training**	**Testing**
Error	39.370%	38.612%

Decision Tree

A decision tree classifies by choosing a threshold on a feature and splits the data according to a 'splitting rule'. Since the features need to be numerical, we had to discard certain features and change how we represented others. For example, we could not convert race into numerical values since this would cause an implicit feature ranking skewing our results. However, education is a feature that can be converted into a numerical value, as a certain level of education can be higher or lower than others in rank. For this reason, we chose only to consider '*age*', '*education*', '*sex*' (This is represented as a binary feature with 1 being male and 0 being female). The tree is then built on the training set and used to predict the binary value of the label (whether or not an individual makes more that $50,000) on the test set.

Decision Tree	**Training**	**Testing**
Error	18.940%	14.778%

The results above are a significant improvement compared to the other methods. This may be because of the binary nature of the dataset, where the prediction is either 1 or a 0, meaning the decision tree (being binary) would have the perfect structure for predicting such a label. Initially it seemed strange that the training error is higher than the testing error, but that can be attributed to the high bias of our dataset. In the dataset, roughly 75% of the entries are that of individuals <= $50,000.

Related Literature

The dataset was taken from UCI's Machine Learning repository (http://archive.ics.uci.edu/ml/datasets/Adult). The dataset has a "Data Set Description" link that contains a list of 17 algorithms ran on the dataset and the results. One of those algorithms is naive bayes, which we also tried. However, their result of 16.12% error is better than our result of 20.43%. They also omit the entries with unknown values for attributes, as this could skew the predictor. Other than that, they don't mention any omitting of other attributes, which differs from us, where we omit attributes that we felt did not have a good distribution or that we felt we could not represent appropriately. This could have lead to the difference in results.

Out of the 17 algorithms ran, NBTree was the best with an error rate of 14.01%. The NBTree is described in Ron Kohavi's paper "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid" (http://robotics.stanford.edu/~ronnyk/nbtree.pdf). Kohavi explains that the NBTree is a hybrid of naive bayes and a decision tree and is most suitable for scenarios where many attributes are significant in predicting the label, but they arent all necessarily conditionally independent. The fact that the attributes weren't completely conditionally dependent or independent could explain the underperformance of both our naive bayes and our decision tree algorithm. A combination of the two in Kohavi's NBTree is the best of both worlds and yields better results in this situation.

Results

The first step we took was to visualize the distribution of each feature and its effect on the likelihood of earning more than $50,000 per year. From our analysis, we concluded that the most useful features for prediction were '*age*', *'education', 'hours per week', 'occupation',* and *'sex'*

Model Recap

k Nearest Neighbors

We attempted to use the k-Nearest Neighbors algorithm to predict if an individual would earn more than $50,000 per year. To calculate the distance between individuals, we looked at the differences of *'age'* and *'hours per week.'* For non numerical features, we added the number of differences between *'education', 'occupation', and 'sex.'*

> *Ex. Consider a1, and a2: Distance = abs(a1.age - a2.age) + abs(a1.hrsPerWeek - a2.hrsPerWeek) + a1.occupation != a2.occupation + a1.sex != a2.sex + a1.education != a2.education*

We did not obtain results since the algorithm failed to finish in a reasonable amount of time.

Naïve Bayes

Worked rather well but is not consistently accurate because the features in the data set may not be completely independent. Grouping of feature keys tended to increase the amount of error.

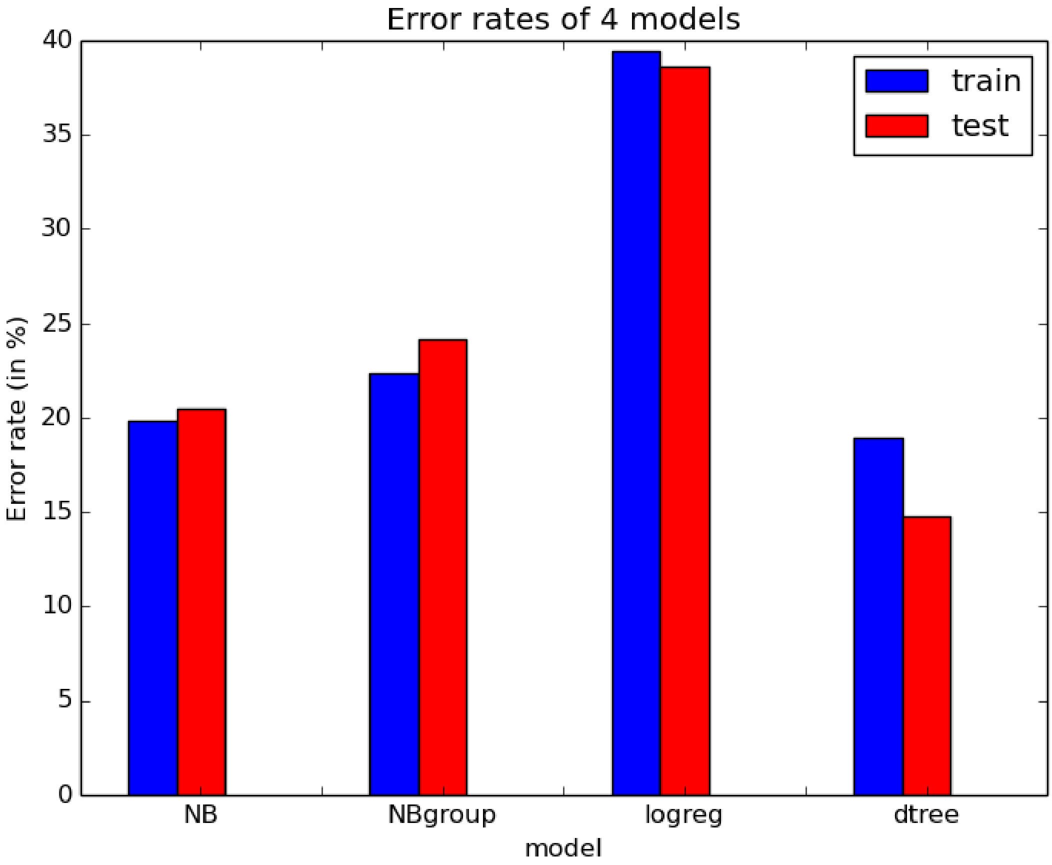
Logistic Regression

This model was the least successful model as the data had a fatal flaw causing incompatibility with the way the classifier works. The binary data points make for poor training in the secondary Y vector and also has issues in not being totally linearly separable.

Decision Tree

This model worked the best out of all of our models giving the highest accuracy. The binary labeled data set was incredibly convenient to integrate with the Decision Tree structure.

Classifier	Train Error	Test Error
Baselines	24.008%	23.623%
Naive Bayes	19.893%	20.432%
Naive Bayes (Grouped)	22.353%	24.128%
Logistic Regression	39.370%	38.612%
Decision Tree	18.940%	14.778%


Error rates of 4 models
40
35
30
25
20
15
10
5
0
Error rate (in %)
train
test
NB
NBgroup
logreg
dtree
model