IDS 561 Final Report:

Census income Analysis to predict income over $60,000

Prathamesh Bapat - Joshua Pollack - Lina Quiceno

Problem Setting

The problem we are looking to solve is predicting the income of individuals based on certain characteristics obtained from the 1990 census data. We will be looking at different parameters and creating models that can best predict this outcome. We turned our problem into a classification question where we will predict if income is less or greater than $60,000. We thought this was crucial for our dataset as we wanted to see what factors were leading to the higher income. As a group we believed that this data can be used for a variety of reasons from individuals to organizations and governments around the country. Being able to understand that variables that directly lead to higher income can help others attain a desired level of professional acumen. While we understand that some of these factors are unchangeable, many of them below can be improved.

It is important to note for our dataset we are looking at income over $60,000 but this is for 1990. Inflation has caused this figure to mean individuals earning over $127,000 in 2021 dollars. The goal of this report is to really demonstrate what factors lead to high earning individuals.

Data Description

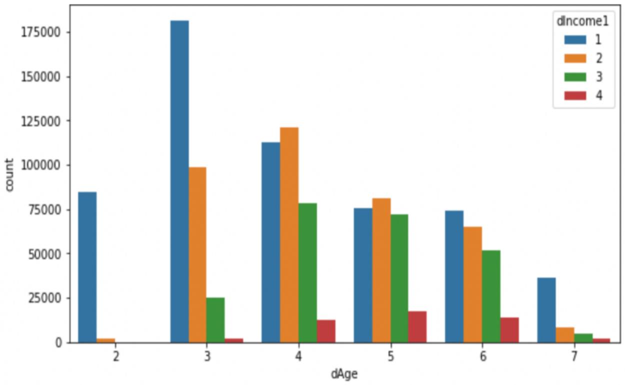
The original authors of the data we got to make our prediction model took the 1990 census data and after evaluating the attributes, they decided to drop the less useful attributes and ended up working with 68 attributes. After this, they collapsed some attributes and discretized the continuous ones.

It is important to highlight that the dataset we are working with is a one percent sample of the Public Use Microdata Samples (PUMS) person records. That is why the authors randomized the data before selecting the 2458285 values we are going to work with.

Our variable of interest is called discIncome1. Through a sql function the authors collapsed the column data into categories from 1 to 4 according to the salary level. As our goal is to predict the income of individuals for over $60.000 we will re-encode this variable as a dummy variable where 1 is above $60.000 and zero below.

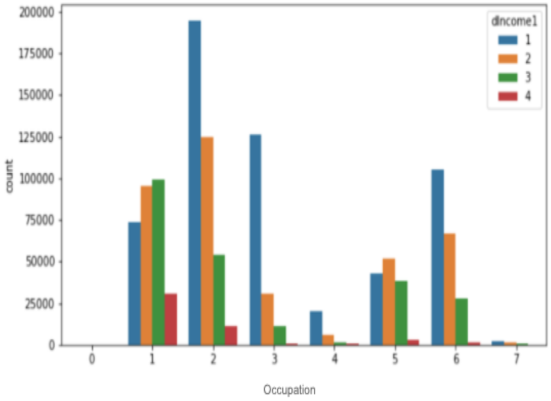
Income was not the only data category that was collapsed into categorical data. Every variable we worked with was changed to better create prediction models. Some of the most interesting changes were taking the ancestry of an individual that was numbered 1 to 1000, the data was collapsed to 1-10 based on different sub categories of possible ancestry.

A large part of our data exploration was looking at how many of the variables were correlated with our label. We looked at every single variable and its relation with the label as bar charts to see what could be analyzed.



As you see here, we included the age group in the X-axis of the chart, and the income level differentiating inside the columns. It became very evident from looking at this that one you hit that age of 40(dAge=4) there was a substantial increase in the amount of people earning over $60,000. This stayed true with an overall shift in the distribution of income at the age of 40. The groups continued to earn more and more as age went on until dAge=7 when many people started to exit the workforce. When working with age we decided to delete all people under the age of 13 from our dataset as that was how the data was already categorized. With labor laws not allowing individuals to work until the age of 16, we felt that keeping that data would not show any insight as all of those individuals might register in other factors such as ancestry and disability but have zero income no matter the other factors. This was not the only distribution we wanted to analyze in determining what categories were important to keep in our dataset.

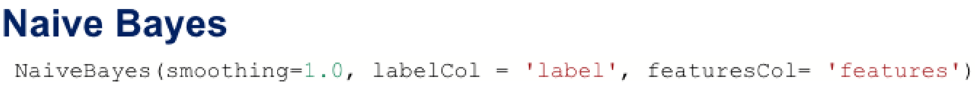
To the left, you will see a distribution of income based on the occupation category that the respondent was included in. This was another variable we turned into categorical because of the amount of options there were and the natural ability to categorize into different sections from the original census data.



Techniques

For our project we used a wide variety of machine learning techniques through pyspark that enables us to create models for prediction. We used Decision Trees, Logistic Regression, Random Forests, Support Vector Machines and Naive Bayes.

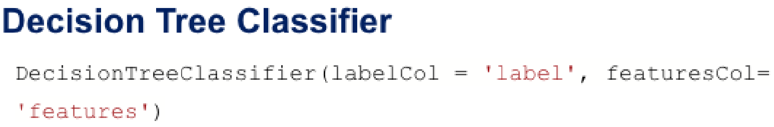
Naive Bayes: a classification model that assumes independence between features and computes the probability of the features applying the Naives Bayes algorithm. In our model we applied the parameter smoothing = 1.0



Naive Bayes

```
NaiveBayes(smoothing=1.0, labelCol = 'label', featuresCol= 'features')
```

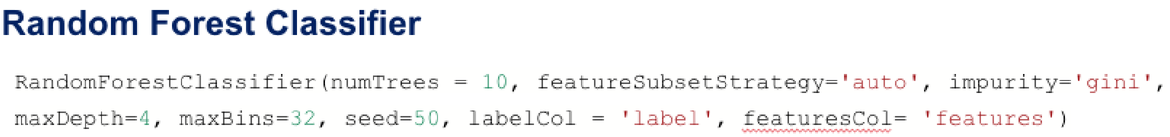
Decision Tree: a greedy algorithm for classification. The algorithm chooses the best splits for the classification of the different labels to optimize the gain in each node. As a measure of impurity the nodes can use either Gini impurity or Entropy for classification problems.



Decision Tree Classifier

```
DecisionTreeClassifier(labelCol = 'label', featuresCol=
'features')
```

Random Forest: is a combination of multiple trees. The advantage of using random forest is the minimization of the possibility of overfitting the model. In the random forest classifier we used the parameter Gini to measure the impurity of the nodes and we also set the number of trees and depth.



Random Forest Classifier

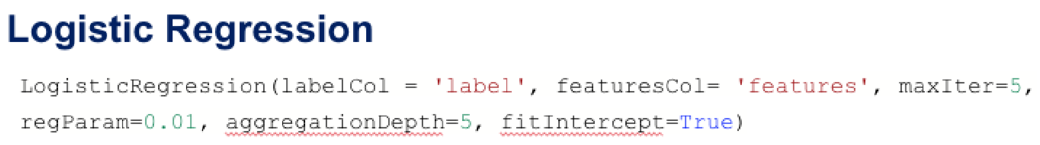
```
RandomForestClassifier(numTrees = 10, featureSubsetStrategy='auto', impurity='gini',
maxDepth=4, maxBins=32, seed=50, labelCol = 'label', featuresCol= 'features')
```

Logistic regression: is a linear model that maps the probability of occurrences into 0 and 1 values and is widely used for classification problems. The formula that use the logistic regression is:

```
f(z)= 1 / 1+e−z
```

We ran the model tuning the following parameters:

- **ReParam:** This is the regularization parameter λ which defines the trade-off between the minimization of the loss and model complexity. Thanks to this parameter we reduce the loss and also avoid overfitting.
- **Aggregation Depth:** suggested depth for treeAggregate (>= 2)
- **fitIntercept:** so the model fits an intercept term.



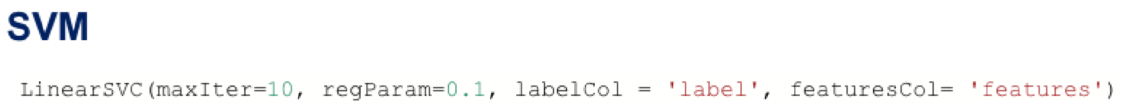
Logistic Regression

```
LogisticRegression(labelCol = 'label', featuresCol= 'features', maxIter=5,
regParam=0.01, aggregationDepth=5, fitIntercept=True)
```

Support Vector Machines
This is also a linear model used for classification data which use the hinge loss.

$$L(\mathbf{w};\mathbf{x},y):=\max\{0,1-y\ \mathbf{w}T\ \mathbf{x}\}.$$

We used the function LinearSVC to run our SVM model which uses a L2 regularized to avoid overfitting. Hyperparameters used in SVM regularization (ReParam) where λ defines the trade off between getting the wider margins possible and the error and maxIter which are the number of iterations for the quadratic programming problem.



SVM

```
LinearSVC(maxIter=10, regParam=0.1, labelCol = 'label', featuresCol= 'features')
```

Results

	ROC AUC		PR AUC	
MODEL	Before	After	Before	After
Logistic Regression	0.985	0.985	0.65	0.75
Decision Tree	0.997	0.999	0.988	0.995
Support Vector Machine	0.904	0.895	0.156	0.181
Naive Bayes	0.662	0.529	0.027	0.038
Random Forest	0.999	0.999	0.983	0.979

We documented the results before taking out the zeros values of income from the dataset and after taking them out, so we could compare how doing this affects our measurement values. As our data was imbalanced we decided to choose the PR curve as a key measurement to evaluate our models. In the table above we can see how the majority of models improve when we remove the zero values (people without a salary) for our income variable.

Feature Importance

Looking into adding interpretability into our model we ran feature importance. This would allow the final users of our model to not only predict the salary of new people that they add to the dataset in the future but also to make decisions and influence in public policies that make an impact in the US population.

It was not surprising that our findings show that the attributes that are more important to predict a high income are Years of education, hours worked the previous year, occupation and Gender. It is also important that the government consider this in the future so its possible to have more information about these sensitive variables.

Role of Team Members

We met to discuss how to approach the dataset and discussed the variable meaninging while also doing the coding. As a team we decided how to treat each variable and what was the business problem and the insights we wanted to tackle. In a second step we work together into the building of the presentation and finally the report in which we all work together.