INTM 551
FINAL PROJECT

ANNUAL INCOME PREDICTION

Enrique Garcia de Vicente
Rafael Cortes Beringola

DESCRIPTION OF THE DATA SET

The Data Set selected for this project belongs to the R resources available at Blackboard. More concretely, in the section of UCI MACHINE LEARNING Data Sets, the chosen set is the second called Adult.

This Data Set has a total of 48842 different facts spread all of them in 14 different attributes. These facts are written either as integers or categorical values. We are going to use this Data Set to predict whether personal income exceeds or not $50K per year depending on the values of these 14 attributes. The extraction of this Data was done by Barry Becker in 1994 from the Census Database.

The 14 attributes where: Age, Working Class, Sampling Weight, Education, Number of years of education, Marital Status, Current Occupation, Relationship, Race, Sex, Capital Gain, Capital Loss, Working Hours per Week and Native Country. For obvious reasons, not all of the attributes contribute the same way in the different models. That is why after creating the model in question, we have seen which variables influenced the most on the outcome, and we have created a new model with the most significant model attributes, reducing computational cost and accuracy.

This Data set also contained missing values so some preprocessing was needed. In first place, some attributes had missing values. Therefore, we first had to identify them and transform them into NA values in order to eliminate the whole row to which it belonged.

Secondly, 8 out of the 14 attributes in the Data Set are categorical factors. This means that, in some cases, we needed to convert 8 attributes to dummy variables. Thus, the number of variables increased from 14 to 93, proving the necessity to eliminate all of them which were not significant and could hinder the model.

In third place, some extra pre-processing has been needed depending on the model implemented. This will be explained in the corresponding method section.

Data Set Link: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

NAÏVE BAYES

The first method that we are going to use to predict is the Naïve Bayes Classifier. Which is a data driven method that makes no assumptions about the data. This method firstly determines which records have the same predictor profile. Secondly, it registers the classes of these records and determines the one that appears most. The new record will be classified as belonging to that class.

The Naïve Bayes classifier requires categorical variables. For this reason, the first step is to convert all the numerical variables to categorical with the only exception of the Sampling Weight, the third column, that is discarded for this analysis. This is because all the numbers that appear in it are different and if applied, it would reduce the accuracy of the model.

After that the data is divided in training, that is used to generate the model, and validation, used to examine the model. The accuracy of the model is using the training data is 83.68 %. For the validation data the accuracy is 82.16 %. Finally, we use the model to predict the class of a new value, that is classified as less than 50K.

LOGISTIC REGRESSION ANALYSIS

As it has been seen in class, the Logistic Regression model performs well when the outcome comes to be a categorical response. In this case, our response can be understood as binary having a personal income higher than $50K (that would be translated into 1) or lower, in which case our variable would be 0. For instance, in order to elaborate our model, we have created a new variable which basically acquires the value of 1 if income is over $50k and 0 in the contrary.

Moreover, we have trained the model as usual with training set so as to evaluate performance of the model with validation set. It is worth saying when creating the model we decided to augment the number of iterations made by the glm() function (by default it does 25 iterations) up to 500 iterations to see if this could enhance model accuracy.

Once elaborating the model with the 93 variables we decided drop all variables which had less than 2.5 Z-Values. This allowed us to bring down the number of variables from 93 to 19 seeing a considerably enhancement in the model accuracy. While accuracy was 0.8232 before, we have

now managed to obtain an accuracy of 0.8448. Finally, as in the previous model we have also predicted that income of the same person, obtaining the same result.

DISCRIMINANT ANALYSIS

The last method we will use to predict class of the earnings of a person will be The Discriminant Analysis. The purpose of this method is to classify a new record into one of the classes by calculating the Statistical/Mahalanobis distance to the centroid of each of them. The model will classify the record as belonging to the closest class.

The model needs to run with numerical variables. For this purpose, we convert some of the categorial variables to numerical (If the race is white or not, which is the most relevant for this analysis case, and the sex). Then we select and combine these new variables and the numerical ones to generate the model.

Finally, we generate the results for our data and the probabilities of belonging to each class. An example of a prediction result can be seen in the following picture:

	Classification	Actual	Score...50K	Score..50K	Propensity...50K	Propensity..50K
1	<=50K	<=50K	27.67	27.40	0.57	0.43

To use the model to predict the class of a new value we must use the coefficients calculated in the model for each class:

$functions		
	<=50K	>50K
constant	-22.8058181	-31.9195111
X9White	6.5925305	6.9302398
X10Male	1.6603738	2.6955889
X1	0.2404751	0.2845386
X3	0.0000226	0.0000232
X5	1.6809002	2.0205516
X11	-0.0000534	0.0000125
X12	-0.0004231	0.0004003
X13	0.2338302	0.2626505

For the sample that we have chosen, the values are:

- <=50K = 27.2749
- >50K = 26.60088

In this case the values are very close, but as the first one is higher, it would be classified as a person who earns less than 50 K.

CONCLUSSIONS

As mentioned before, we have predicted income for the same sample using the three models. Our sample was:

Age	***Work Class***	***Sample Weight***	***Education***	***Education Hours***	***Marital Status***	***Occupation***	***Relationship***
39	Private	215646	Assoc-voc	11	Married	Exec/Manager	Husband

Race	***Sex***	***Capital Gain***	***Capital Loss***	***Hours/Week***	***Native Country***
White	Male	5000	0	40	USA

For the two of the models: Naïve Bayes and Logistic Regression the prediction is that the income is higher than $ 50K. On the other hand, Discriminant Analysis predicts it to be lower than $ 50K, but with very close values on both sides. These discrepancies are due to the fact that the data that we have chosen is in the middle ground, that can not be clearly classified in one of the classes.

Last but not least below we can see a table regarding model's accuracies:

Model	***Naïve Bayes***	***Logistic Regression***	***Discriminant Analysis***
Accuracy	82.16%	84.48%	80.67%

From the results, we can see how Logistic Regression is the one which fits better with this Data Set. Even though accuracies could be higher, they depend on the kind of Set and features such as their facts quality, dispersion and variance among others. In this case, the three models perform at a similar level, even though each model is more-commonly used for certain Data Set. For instance, Naïve Bayes is recommended when Data Sets have categorical variables, Logistic Regression when the output is a categorical value and Discriminant Analysis when the Data follows a normal distribution. For sure, Naïve Bayes should work adequately because the data includes several categorical variables.

Nevertheless, this data has a higher complexity because it has many categorical variables that have to be converted into binary for some models, due to this, in some models, we have reached the number of 92 variables. This, along with the fact that our data does not follow a completely normal distribution and due to the quality of the data, explain the fact that the accuracies are not higher. But as all of them are very similar, we believe that we have reached a good performance for the three models.