IV. BINARY PREDICTION OF INCOME

UNIVARIATE STATISTICS

Our data are from the census bureau database. Income is our outcome variable. It has been dichotomized into income greater than or equal to $50,000 and income less than $50,000. In our dataset of 32,561 individuals, 76% have an income less than $50,000. This means that if our goal is prediction of income, our null rate upon which we must improve is 76%.

Our predictors[1] consist of age, type of work (workclass), highest level of education attained (edu), marital status, occupation, relationship, race, sex, capital gains recorded (capital_gain), capital losses recorded (capital_loss), hours worked per week (hours_per_week), and native_country. Table 4A shows five-number summaries for the numeric variables.

TABLE 4A. FIVE NUMBER SUMMARIES OF NUMERIC PREDICTORS.

	Mean	Median	Std. Dev.	Min	Max
Age	38.6	37	13.6	17	90
Capital Gain	1077.6	0	7385.3	0	99999
Capital Loss	87.3	0	403.0	0	4356
Hours per Week	40.4	40	12.3	1	99

Note that age is positively skewed; both capital gain and capital loss suffer from heavy zero-inflation; and hours worked per week is leptokurtic, suffering from large peakedness.

Capital Gain has a handful of observations with extremely large values suggesting an outlier group; however this group appears to be a legitimate part of the sample and having such an extreme value for Capital Gain may have prediction value for income (e.g. it may be easier to have large Capital Gain if your Income is also high).

[1] The variable fnlwgt, the number of people the census takers believed a certain observation represented, was removed to simplify the analysis. The variable edu_num, the number of years of education, is simply the edu variable in numerical form; it was removed as well.

CHART 4A. TYPE OF WORK.

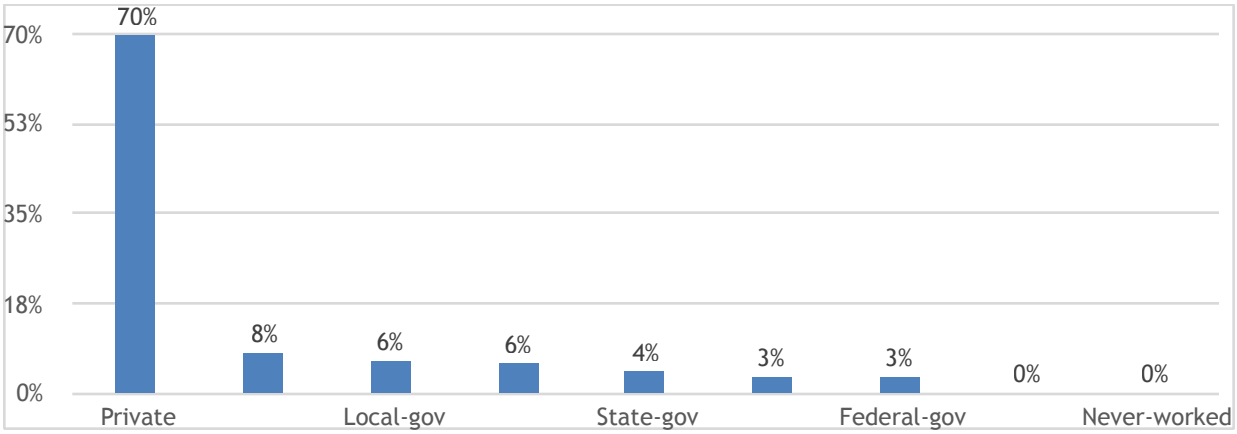


CHART 4B. LEVEL OF EDUCATION ATTAINED

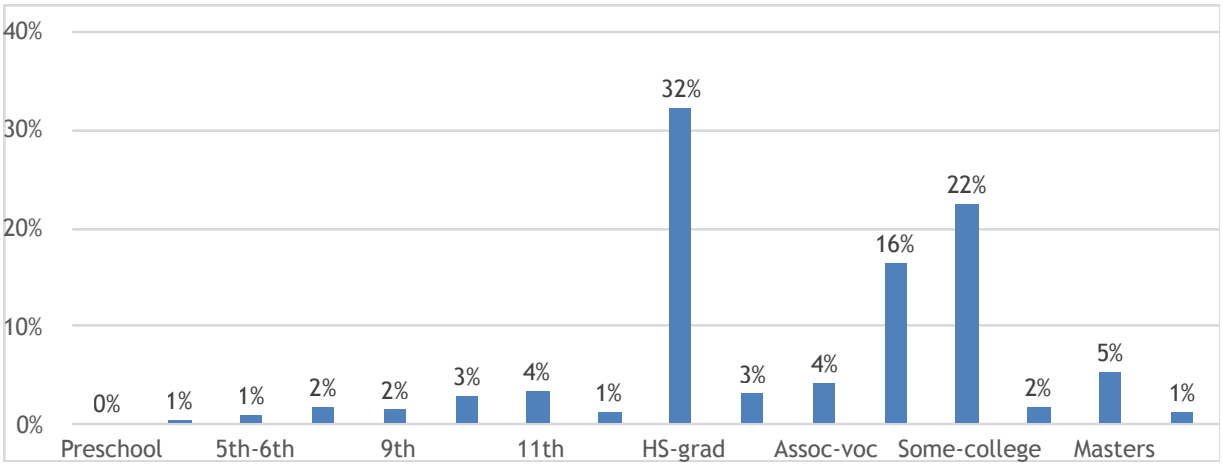


CHART 4C. MARITAL STATUS

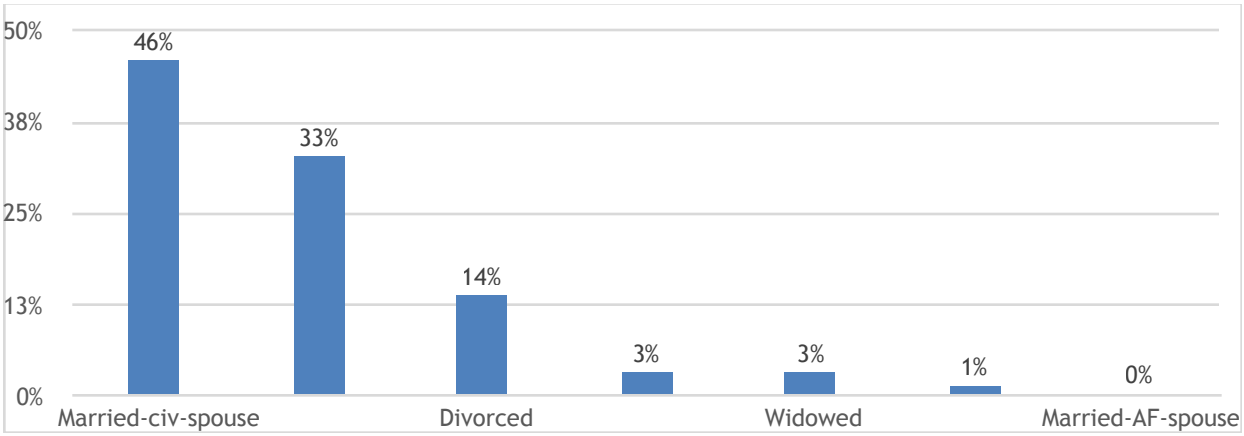


CHART 4D. OCCUPATION.

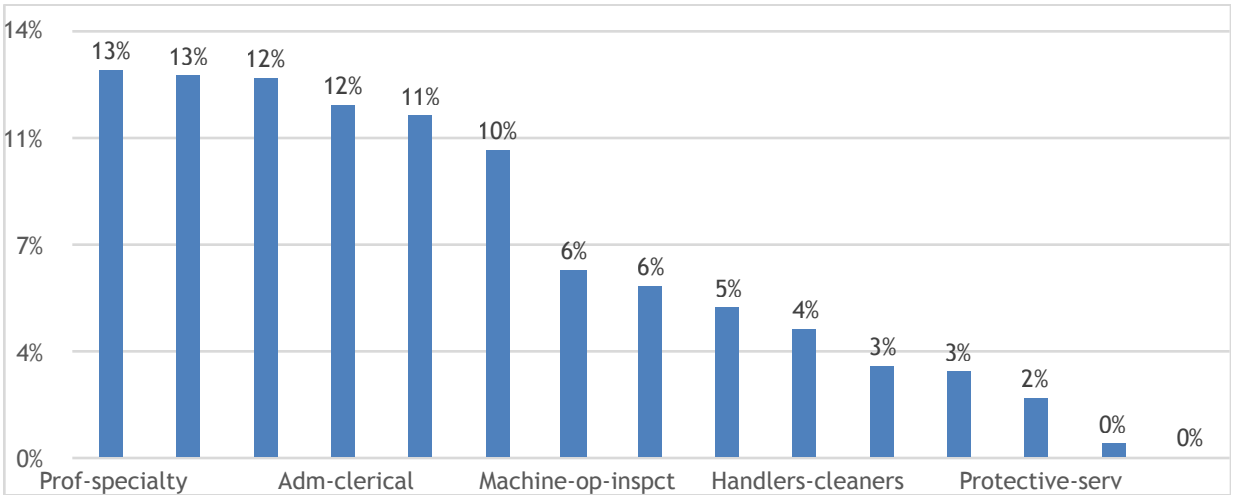
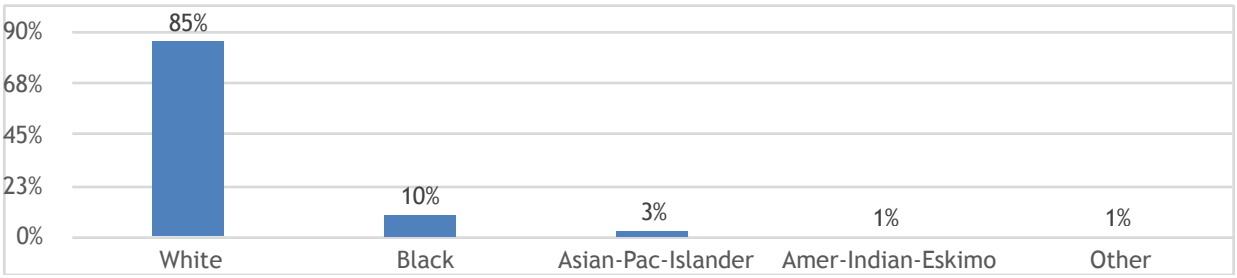


CHART 4F. RACE.



Our sample was about 33% females and 67% males. A total of 41 countries were represented with the majority of observations coming from those with origin in the US; about 2% of the sample originated from Mexico and 1.8% were unknown.

Since our outcome variable, income, is dichotomous, we are looking at some sort of classification technique. We shall consider two classification methods: logistic regression (LR) and random forest (RF). We will perform model selection within each method (looking at fit and predictive power) and then compare the best model of each method in terms of predictive power to make the final selection.

I chose to use these two methods because they are antithetical in their strengths and weaknesses but complementary when considered together since they cover a wide variety of possible classification methods that exist out there. Logistic regression is essentially a linear model (predictors and the logit of the response are linearly related) and if the true relationship between the logit of income and the other predictors is linear and fairly simple, logistic regression provides a rich interpretation to how the predictors and the response are related. If this relationship is actually highly nonlinear and complex, the hyperplane fit by the prediction equation in a linear model will capture the signal in the data poorly.

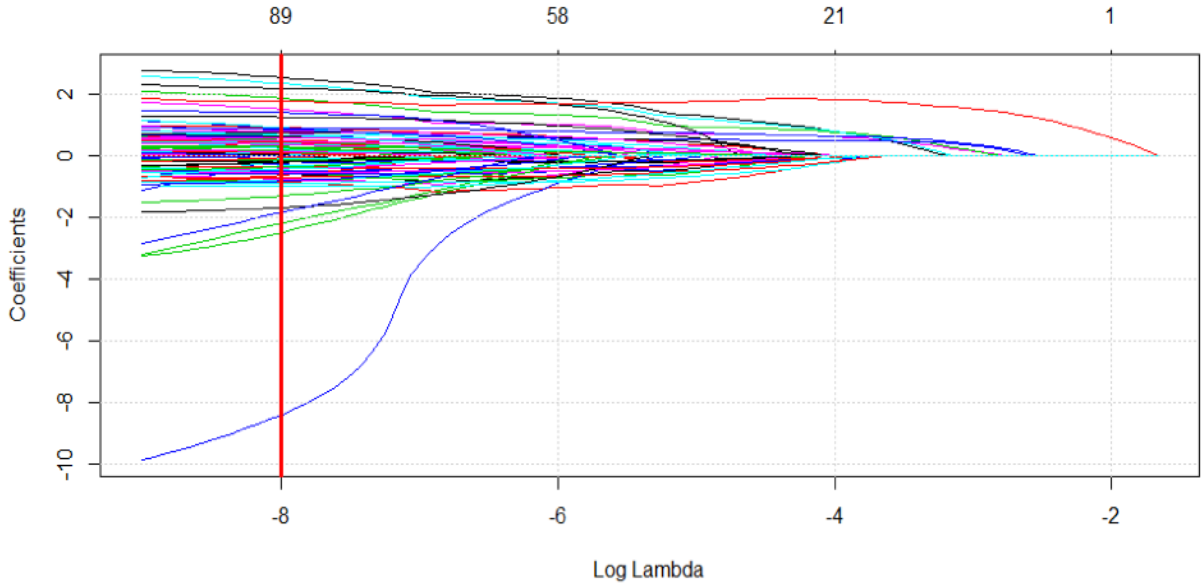
Instead we can turn to a piece-wise global model that partitions the predictor space into hyperrectangles to which we can assign some constant prediction i.e. a decision tree. Such a method is highly sensitive to the data and will overfit spectacularly unless we use some form of ensemble learning where we grow a whole forest of decision trees and take a majority vote from them. Even better would be to decorrelate the decision trees so that they can more easily capture every single predictor and interaction needed to optimize prediction. The result is a random forest, our second method. Where it fails in interpretation (as a black box method), it excels in predictive power. These two methods represent the majority of classification methods out there in the sense that they highlight the dichotomy between powerful black box methods and interpretable linear models.

LOGISTIC REGRESSION

Given the large ($p = 12$) number of predictors, we need to use some sort of model selection process to weed out the predictors that do not significantly contribute to predicting Income. If there were fewer predictors we could take a more involved approach and use G^2 and ΔG^2 to select a good-fitting model. Instead, I will elect to use some sort of regularization. Our exploratory data analysis seems to tell the story that not all predictors are strongly related to Income, some not even weakly related. This would suggest Lasso regularization so that the predictors that aren't related can be removed.

We fit a Lasso glm and found that the cross-validated lambda was $\lambda = 0.0002$ which is basically the same as an unpenalized glm. As you can see below in Graph 4A, at this value of lambda, barely any variables are shrunken to zero; in fact only seven are shrunk to zero and they are mostly dummy variables of the country factor. These results indicate that the 12 predictors are all important in predicting income and the best model we can run is a regular logistic regression with all predictors included.

GRAPH 4A. VARIABLE SELECTION AS A FUNCTION OF LAMBDA TUNING PARAMETER.



At the value of lambda selected through cross-validation (the vertical red line) very few variables have been shrunken to zero which is the whole point of the lass

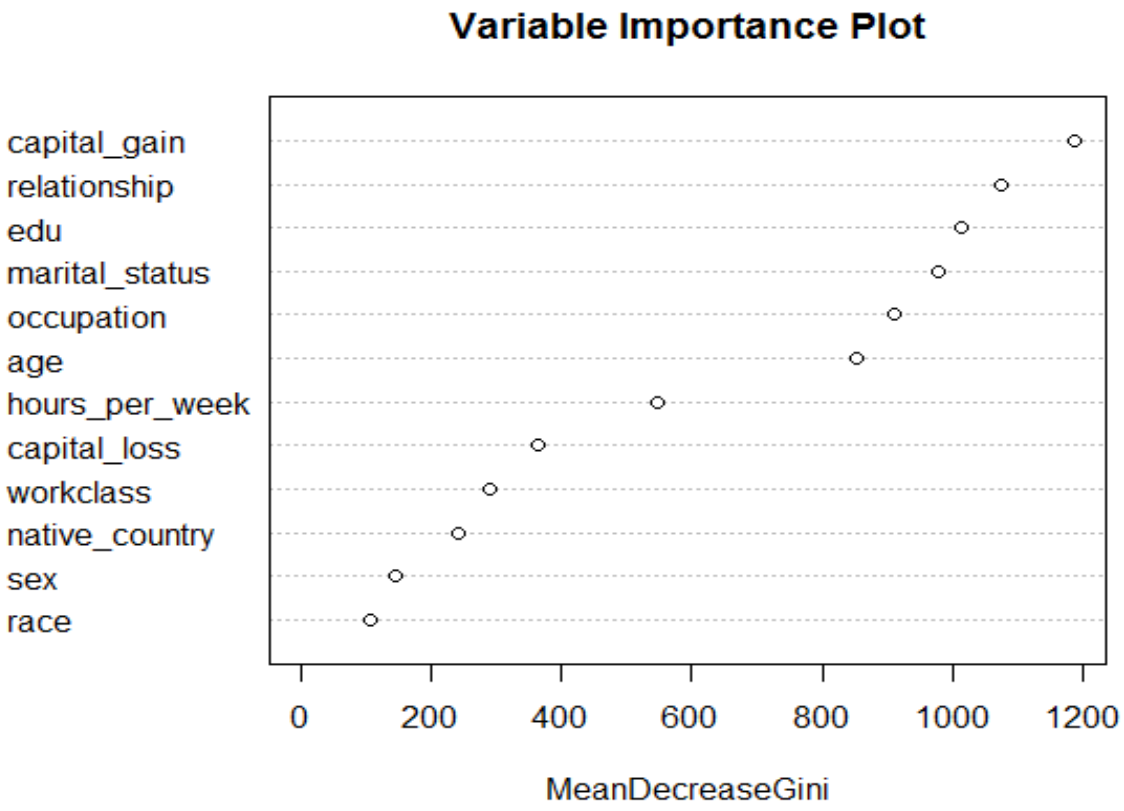
If our model fits the data well, then it's deviance is chi-squared with degrees of freedom 32463. We can use this to test the absolute fit of our model to the data as we expect our deviance divided by its degrees of freedom to be about 1. For our model, $G^2_{full}/df_{full} = 0.63$. which indicates our deviance is fairly small and that our model fits the data well.

RANDOM FOREST

The tuning parameter for random forests is $mtry$, the number of randomly selected predictors to use at each split in each tree. A common heuristic is the square root of the number of predictors but we can actually tune $mtry$ by a less computationally-intense analog of cross-validation in random forest models called out-of-bag error (OOBE) estimates. Using this methodology, we find that $mtry = 2$ minimizes the OOBE.

As previously mentioned there isn't much interpretation that can be extracted from a random forest model. We can, at best, extract a variable importance plot which tells us, on average (across trees) how much a variable decreases Gini node purity (roughly, the variance of the class probabilities in an internal node). Graph 4B below shows that Capital Gain was the most important predictor of Income followed closely by Relationship, Education, Marital Status, Occupation, and Age.

GRAPH 4B. VARIABLE IMPORTANCE PLOT AS MEASURED BY GINI NODE PURITY.



COMPARISON OF METHODS

To compare predictive accuracy, we want to look at the test error rate, which takes the form of the misclassification rate with our discrete outcome. We can examine the error rate a bit closer by looking at conditional probabilities in the confusion matrix - specifically the true positive rate and true negative rate of the procedure. For example, of those who did have incomes above 50k, how many were correctly predicted by the statistical method? For those who had incomes below 50k, how many were correctly predicted?

Misclassification Rate

For the logistic regression, the 5-fold cross-validated misclassification rate is 14.89%. For the random forest, the out-of-bag estimate of error rate is 17.65%, which should be equivalent to the misclassification rate estimate if we had used cross-validation.

Cross-Validated/OOB Confusion Matrix

		LR Predicted	
		<=50K	>50K
Actual	<=50K	23002	1717
	>50K	3130	4711

The logistic regression has a cross-validated true positive rate (prediction of >50K when observation is >50K) of 93.1% and a cross-validated true negative rate (prediction of <=50K when observation is <=50K) of 60%.

		RF Predicted	
		<=50K	>50K
Actual	<=50K	24665	55
	>50K	5693	2148

The random forest has an OOB true positive rate (prediction of >50K when observation is >50K) of 99.8% and an OOB true negative rate (prediction of <=50K when observation is <=50K) of 27.4%.

CONCLUSIONS

In terms of overall misclassification rate, logistic regression outperforms the random forest with 2.76 percentage points more of precision; both perform the null classification rate by about 10%. Logistic regression really outperforms the random forest when we examine their confusion matrices. Both methods perform excellently in terms of predicting the majority class (<=50K) but differ wildly in how well they predict the minority class (>=50K). Logistic regression correctly predicts 60% of those earning more than $50,000 whilst the random forest correctly predicts only 27.4%. Perhaps this poor performance for the random forest is a sacrifice for the high precision in predicting those who earn less than $50,000 (99.8% accuracy compared to 93.1% in logistic regression). In terms of balanced performance logistic regression is the champion. The only situation in which random forest would be preferred would be if correct classification of (<=50K) was much, much more important than correct classification of (>=50K). Then we could justify the poor classification of (>=50K) if it meant the near perfect classification of (<=50K). In this sense, the random forest is a huge improvement on the null rate in that it keeps the perfect classification of (<=50K) of the null rate whilst raising the 0% classification of (>=50K) of the null rate up to 27.4%

What does the superiority of logistic regression tell us about our data? Our data appears to exhibit a somewhat linear (in the logit of income) and simple structure, not requiring interactions and polynomial terms that random forest would have been able to approximate. Not only does the logistic regression fit the data very well but it has the added benefit of containing easily interpretable coefficients. We know that the 12 predictors all significantly predict odds of making greater than $50,000 but now we can get a clearer picture of the manner in which they influence the odds, whether it be the difference in odds between those originating from the US and those originating from Mexico or the increase in the odds as age increases.

R CODE

```
require(randomForest)

adult <- read.table("https://archive.ics.uci.edu/ml/machine-learning-
databases/adult/adult.data",sep=",")
names(adult) <- c("age", "workclass", "fnlwgt", "edu", "edu_num",
"marital_status", "occupation", "relationship", "race", "sex",
"capital_gain", "capital_loss", "hours_per_week", "native_country",
"income")
adult[["edu_num"]] <- NULL
adult[["fnlwgt"]] <- NULL




#EXPLORATORY DATA ANALYSIS------------------
boxplot(adult$age,ylab="Age",main="Boxplot of Age")
with(adult,plot(age,income))

#age and income
hist(adult[adult$income==" <=50K","age"],
col=rgb(1,0,0,0.5),xlim=c(15,100), ylim=c(0,4000), main="Age and
Income", xlab="Age")
hist(adult[adult$income!=" <=50K","age"], col=rgb(0,0,1,0.5), add=T)
legend(x="topright", c("Income less than 50k","Greater than 50k"),
lty=c(1,1), lwd=c(10,10),col=c(rgb(1,0,0,0.5),rgb(0,0,1,0.5)))



#PENALIZED LOGISTIC REGRESSION---------------
require(glmnet)
x <- model.matrix(income~.,data=adult)[,-1]
glmmod <- glmnet(x,y=adult[["income"]],alpha=1,family="binomial")

plot(glmmod,xvar="lambda")
grid()
abline(v=-8,col="red",lwd=3)

cv.glmmod<-cv.glmnet(x,y=as.numeric(adult[["income"]])-1,alpha=1)
plot(cv.glmmod,ylab="Misclassification Rate")
cv.glmmod$lambda.min #cross-validated lambda is almost zero
```

```
sum(glmmod[["beta"]][,77]==0) #7 coefficients shrunk to zero

#ORDINARY LOGISTIC REGRESSION---------------

fit.log <- glm(formula = income ~ .,data = adult, family =
binomial(link = 'logit'))

fit.log[["deviance"]]/32463


#RANDOM FOREST------------------------------
require(randomForest)
require(caret)
set.seed(123)

bestmtry <- tuneRF(adult[,-13],adult[,13], ntreeTry=100,
stepFactor=1.5,improve=0.01, trace=TRUE, plot=TRUE, dobest=FALSE)

fit.rf <- randomForest(income ~ ., data=adult, mtry=2,
importance=TRUE, do.trace=100)

varImpPlot(fit.rf,type=1)

#COMPARISON---------------------------------

#misclassification error rates
fit.rf
#confusion matrices
fit.rf[["confusion"]]




#5-fold cross-validation for logistic regression confusion matrix
set.seed(12)
adult_nh<-adult[-which(adult$native_country==" Holand-Netherlands"),]

folds_i <- sample(rep(1:5, length.out = nrow(adult_nh))) #assign
observations to folds
for (k in 1:5) {
  test_index <- which(folds_i == k)
  train_fold <- adult_nh[-test_index, ]
```

```
  test_fold <- adult_nh[test_index, ]

  fit.log <- glm(formula = income ~ .,data = train_fold, family =
binomial(link = 'logit'))
  y <- test_fold[["income"]]!=" <=50K"
  yhat<-round(predict(fit.log,newdata=test_fold,type="response"),0)
  assign(paste("confusion",k,sep="_"),table(y,yhat))
}

log.confusion=confusion_1+confusion_2+confusion_3+confusion_4+confusio
n_5
```