STA 141C Final Project (Winter 2020)

Author: Eva Chen, Muhammad Ali, Yaosong Zhan, Sukhween Bhullar

Introduction

Goal: The goal of the following project is to analyze a Census dataset in which we classify the income group, >50K (greater than $50,000) or <=50K (less than $50,000), of a specific individual. This data will be split into two parts: training and testing. We intend to implement Bag of Little Bootstraps with the Generalized Logistic Model on the training data in hopes of predicting the income groups of our test data. Since the distribution of our dataset is unknown, Bag of Little Bootstraps would be useful because then we do not have to worry about the underlying assumptions of distributions. Through this method we subsample our data to increase computational efficiency and get better estimates for our model.

Source of Data: The data was acquired from the UC Irvine Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Adult. The data was extracted from the 1994 Census database. It consists of 15 attributes of which 1 is the response variable income (>50K and <=50K).

Questions:

- Does the model accuracy rate change after implementing BLB on the Logistic Regression?
- What is the predicted class of an individual given their demographics?
- Amongst the different BLB Logistic Regression estimation methods we used, which one is the "best"?

Variables:

income (Categorical):

- Levels: >50K, <=50K
- Ex. If someone falls in the >50K group they earn more than $50,000. If someone falls in the <=50K group thy earn less than $50,000

age (continuous):

- Integer value greater than 0
- Minimum: 17, Maximum: 90

workclass (Categorical):

- Variable dropped from Logistic Model
- Working class category that the individual belongs to
- Levels: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked

fnlwgt (Continuous):

- Variable dropped from Logistic Model

- Variable dropped from Random Forest
- The number of people with those specific “credentials”
- Ex. 77516 individuals are age 39, belong to State-gov, with a Bachelor’s degree, were never married, etc and resulted in an income_group greater than 50k
- Minimum: 13769, Maximum: 1484705

education (Categorical):

- Variable dropped from Logistic Model
- Levels: 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: (Continuous)

- Variable dropped from Random Forest
- Number associated with the education category
- Ex. 13 is equal to Bachelors degree
- Minimum: 1, Maximum: 16

marital_status: (Categorical)

- Variable dropped from Logistic Model
- Levels: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, married_spouse-absent, Married-AF-spouse

occupation: (Categorical)

- Variable dropped from Logistic Model
- Levels: 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: (Categorical)

- Variable dropped from Logistic Model
- Levels: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
- Example: If someone’s marital_status is Married-civ-spouse then the relationship will be either Wife or Husband

race: (Categorical)

- Levels: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black

sex: (Categorical)

- Levels: Female, Male

capital_gain: (Continuous)

- Variable dropped from Logistic Model
- Minimum: 0 , Maximum: 99999

capital_loss: (Continuous)

- Variable dropped from Logistic Model

- Minimum: 0, Maximum: 99999

hours_per_week: (Continuous)

- Minimum: 1, Maximum: 99

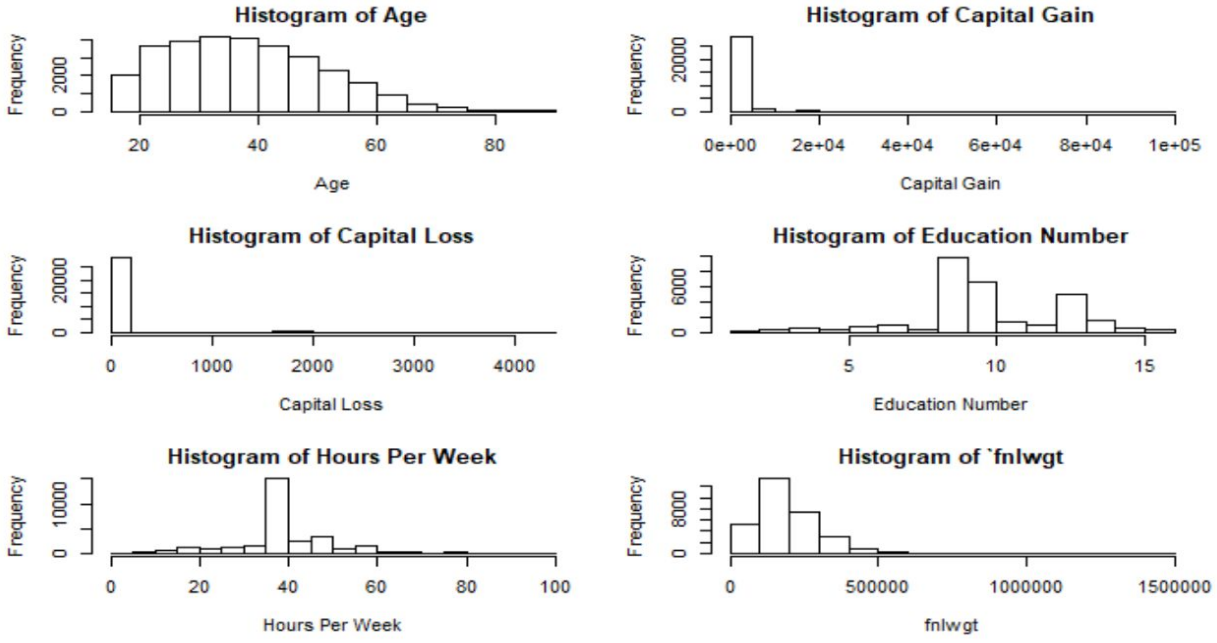
native_country: (Categorical)

- Variable dropped from Logistic Model
- Variable dropped from Random Forest
- Levels: 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.

Exploratory Data Analysis:

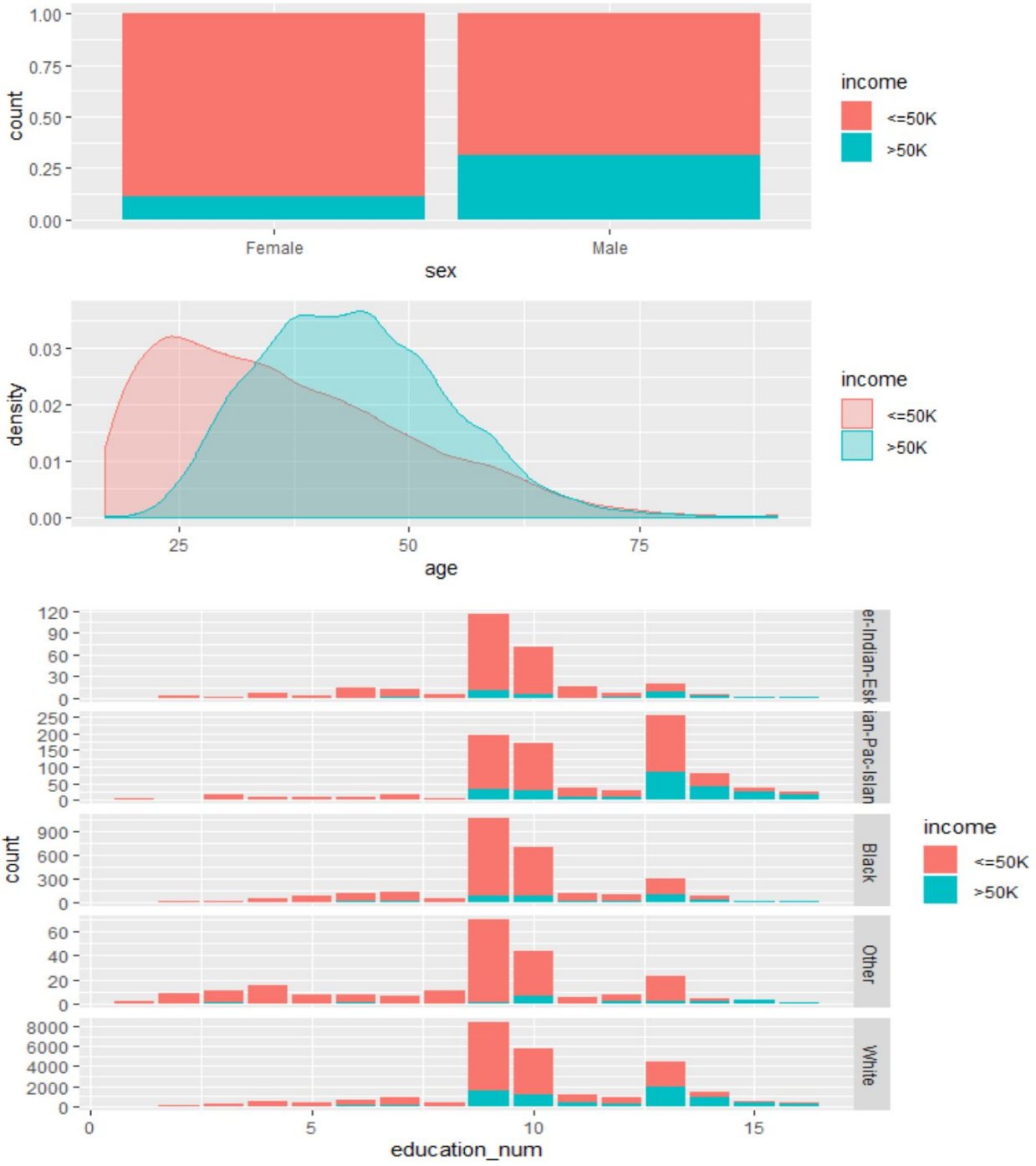
Before we begin our model analysis, we did some general exploratory data analysis to explore variables of interest.

Histograms of all the quantitative variables: age, capital_gain, capital_loss, education_num, hours_per_week, and fnlwgt, were used to visualize the general distributions. From these histograms we may consider taking out the variables 'capital_gain' and 'capital_loss'. The variable 'fnlwgt' can also be dropped because these are weights that are used by the Census, it's not exactly something an individual can enter in. More reasoning behind which variables were dropped will be explained later in the 'Methodology and Process' section.



The following graphs show a distribution of our response variable, income. The first chart compares income distribution between males and females. The second graph compares how

income distribution varies amongst different ages. The second chart shows the distribution of income split across different education_num values (each number equates to an education level, ex. 13 is Bachelor's degree) and races. The third chart shows the distribution of income split across different education_num values (each number equates to an education level, ex. 13 is Bachelor's degree) and races.



From data visualizations, we find some features of our dataset. First, the percentage of income > 50K in the male group is higher than that in the female group. Second, there are few adults

whose incomes are more than 50K in the lower education number groups. On the contrary, in groups with education numbers greater than 15, the percentage of income > 50K is nearly 100%. We find this feature in all race groups and sex groups. Third, we find that the distribution of age in income < 50K group and income > 50K group are different. In the income < 50K group, the peak of age distribution is below 25, while in the income > 50K, the maximum of the distribution is around 40. Last, the relationship between income and hours per week is not monotonous. Sixty hours per week has the highest percentage of income > 50K.

Methodology and Process:

We had to use multiple methods to acquire a model that would fit the structure of our data. We ended up using the following approaches:

- Generalized Logistic Regression
- Bootstrap
- Customized prediction accuracy calculator
- Random Forest

Generalized Logistic Regression

We decided to fit the data using a binomial logistic model because our response variable of interest, income, is binary and our explanatory variables are mixed with some categorical and some continuous. From our data exploration, we can see that the variables capital gain and capital loss consisted of mainly zeros, so it was dropped from our model as it would not provide a substantial explanation for the relationship between subjects and the income group they fall under. We also chose to exclude education because the variable education number will provide the same information in years. After several attempts of randomly sampling the data in preparation of the bootstrapping process, we found that not all levels of certain categorical variables were showing up. For example, the work class variable had a very small amount of subjects that were without pay so this level showed up in some subsamples but not in others. This made estimation difficult since the data frames were not equal. After much deliberation, those variables were cut from the model.

As you can see, we used logistic regression without bootstrapping and included all the variables, the end result was a very long list of coefficients.

```
[1] "The number of estimates in the full logistic model: 96"
```

By removing all the categorical variables minus race and keeping all the quantitative variables minus ‘capital_gain’, ‘capital_loss’ and ‘fnlgwt’ we get a manageable model.

	coeffs_simple
(Intercept)	-9.43209664
age	0.04574728
race Asian-Pac-Islander	0.26940233
race Black	0.13873763
race Other	-0.43341495
race White	0.52948538
sex Male	1.16844287
education_num	0.35169220
hours_per_week	0.03254820

Bootstrap (with Generalized Logistic Model)

We implemented bootstrap to obtain an estimate of what the true coefficients are. We separated the data into ten subsections, and then for each subsection we bootstrapped a thousand times. For every iteration of bootstrap, we fitted a logistic model and placed the extracted coefficients into a dataframe. The data frame consists of ten lists that contains 1000 samples of each coefficient. Subsequently, we took the quantiles of each coefficient and reduced the ten lists to acquire an overall interval. The confidence interval covered a wide range which was not very informative when trying to estimate which income group a subject was in. We tested both the upper bounds and lower bounds with the same subject. The subject was a 37 year old Black male with ten years of education (some-college) that worked 80 hours a week. From the data, we know that this subject makes over 50,000 a year. Using the lower bound coefficients, we found a 16 percent chance that this subject makes over 50,000. When we used the upper bound coefficients we found approximately 75 percent chance that the subject makes over 50,000 a year. The estimates were vastly different so we found our model impractical. We further explored by carrying out the same method with the mean and median, however these gave us similar results to the confidence intervals. Our model had a low accuracy of classifying a subject into the right income group.

After getting running the bootstraps we implemented three different ways to get an estimate of the slopes by using:

- confidence interval
- median
- mean

The confidence intervals give us a lower and upper bound for beta estimates:

	Coefficients_LowerBound	Coefficients_UpperBound
(Intercept)	-11.24286602	-10.28211025
age	0.04343381	0.04856240
raceAsian-Pac-Islander	1.06465610	1.98076055
raceBlack	1.00086413	1.83660399
raceOther	-1.18986858	0.31859666
raceWhite	1.41996309	2.22071075
sexMale	1.09135085	1.26749628
education_num	0.33761108	0.36833569
hours_per_week	0.02976904	0.03593301

The medians give the following beta estimates:

	Coefficients_Median
(Intercept)	-10.73265720
age	0.04599597
raceAsian-Pac-Islander	1.49339778
raceBlack	1.38458182
raceOther	-0.36718162
raceWhite	1.78373416
sexMale	1.17495195
education_num	0.35328806
hours_per_week	0.03276443

The means give the following beta estimates:

	Coefficients_Mean
(Intercept)	-10.74239558
age	0.04598458
raceAsian-Pac-Islander	1.50617095
raceBlack	1.39740493
raceOther	-0.38033655
raceWhite	1.79669622
sexMale	1.17520633
education_num	0.35324032
hours_per_week	0.03277506

Next Steps: Classification Trees and Random Forest

After understanding that the Logistic Regression Model may not be the best fit for our data we understood that we might need to implement another model, the classification/regression tree. Because this model can be used for both categorical and quantitative variables, in the form of a classification tree and a regression tree respectively, it would work perfectly with the kind of data we're using. To at least see a general trend, we ran the full model, but we sadly ran into an error saying that we could not have variables with more than 32 levels. This led us to drop the

variable “native_country” as it had the most levels. We also chose to drop “fnlwgt” because this variable is just an estimation for weights used by the Census and was not applicable to our model. Finally, we removed “education” because we already had a variable called “education_num” that was associated with the education level of an individual, keeping both would’ve been redundant. The next step was to create multiple trees, so we applied a bootstrap method to randomly select columns to create multiple trees, in other words a Random Forest.

The initial tree classification model is used on age, workclass, education, marital_status, occupation, relationship, race, capital_gain, capital_loss, hours_per_week and sex. The next step was to create multiple trees, to create a random forest.

Next Steps: Bag of Little Random Forests

Unfortunately with the time restrictions we only had enough time to create one BLB model, and were not able to create the Bag of Little Random Forests, but to further our knowledge we wanted to at least attempt the code.

To create our Bag of Little Random Forest we would use bootstrap and then implement a Random Forest model. The approach was to resample from each one of the 10 subsamples 1000 times. Then we want to sample eight columns from our final list of eleven columns consisting of age, workclass, education, marital_status, occupation, relationship, race, capital_gain, capital_loss, hours_per_week and sex. In every iteration, we would create our prediction tree off of our training data and then predict the classification probabilities with our testing data. In the end we wanted to create 1000 different classification probabilities for our testing data, resulting in approximately 6,000,000 rows of data to process for 10 different bootstrapped datasets. To visualize, one can imagine one iteration of a test data set consisting of ~6000 rows, and we create 1000 different probabilities. After getting our 1000 probabilities for each row, we want to calculate the mean probabilities of that row being classified as >50K and <=50K. Then we simply compare which probability is higher to classify each row. This would create a mean prediction value for each one of the subsets, so now we go from ~6,000,000 predictions for 10 subsets to ~6,000 mean predictions for 10 subsets, then we want to create 1 final prediction set consisting of ~6,000 rows. The final dataset would’ve been benchmarked for accuracy against the actual test data values.

Customized Prediction Accuracy Calculator

Our team wanted to calculate how accurate our models were, but this proved to be difficult with the structure of our final dataset. In the case in which we don’t use bootstrapping and simply use the glm or tree function once without repetition, we could easily use the confusionMatrix function to get the accuracy of our results by testing the respective models on our testing data. With bootstrapping, we would need to create a method that deals with the complexity of our data structure that have gone through transformations. Through a series of simple if else statements

we were able to create a function for bootstrapping with the logistic regression model. A similar function to calculate the Random Forest accuracy was created. This allowed us to calculate the accuracy of our models. An accuracy calculator returns a vector of the number of:

- ❖ incorrect values categorized as >50K
- ❖ correct values categorized as >50K
- ❖ incorrect values categorized as <=50K
- ❖ correct values categorized as <=50K

We then used these numbers to calculate the accuracy rate, which is the correct/total, which will be shown in the Analysis section.

Analysis and Improvements:

For our final results, we got the coefficients from the Logistic Regression (these coefficients include implementations with and without Bag of Little Bootstraps). The results have been formatted as a table in the following order: Simple Logistic Regression without Bootstrap, then BLBGLM with Mean, BLBGLM with Median, BLBGLM with Confidence Intervals.

	Coefficients_Simple	Coefficients_Mean	Coefficients_Median	Coefficients_LowerBound	Coefficients_UpperBound
(Intercept)	-9.43209664	-10.74239558	-10.73265720	-11.24286602	-10.28211025
age	0.04574728	0.04598458	0.04599597	0.04343381	0.04856240
race Asian-Pac-Islander	0.26940233	1.50617095	1.49339778	1.06465610	1.98076055
race Black	0.13873763	1.39740493	1.38458182	1.00086413	1.83660399
race Other	-0.43341495	-0.38033655	-0.36718162	-1.18986858	0.31859666
race White	0.52948538	1.79669622	1.78373416	1.41996309	2.22071075
sex Male	1.16844287	1.17520633	1.17495195	1.09135085	1.26749628
education_num	0.35169220	0.35324032	0.35328806	0.33761108	0.36833569
hours_per_week	0.03254820	0.03277506	0.03276443	0.02976904	0.03593301

After implementing the accuracy calculator we created a table of accuracy rates (in %) for the mean, median, and confidence intervals for the Bag of Little Bootstraps with the General Logistic Regression Model and joined it with the accuracy rates for the simpler models that don't use Bag of Little Bootstraps. The table input for accuracy is based on the following models: GLM on the full model, GLM without bootstrap (selected), then BLBGLM with mean, BLBGLM with median, BLBGLM with confidence intervals. The following accuracy rates were acquired (in %):

GLM (full, no bootstrap)	Model 1, no bootsrap	Mean Rate	Median Rate	CI Rate (Lower)	CI Rate (Upper)	Tree Rate	Random Forest Rate
84.58223	75.97812	75.96154	74.63528	77.88462	79.7248	83.96883	83.96883

From the above table it is evident that the most accurate model is the General Logistic Regression model that did not use any bootstrapping as it had the highest accuracy rate at 84.58%. On the contrary the GLM regressed against the variables that would eventually be used in the BLB, has a lower accuracy rate at 75.96%. We would expect this because the first model uses all the variables, implying that the smaller model may have dropped some significant variables. Given the circumstances of our approach, we had to drop some variables to make this BLBGLM possible, even though a BLBGLM using all variables would be amazing it is not practical in the sense that it creates 96 different estimates. So to move forward we applied our BLB to a logistic regression model that has less variables. Thankfully, the numbers showed that BLBGLM was effective since the accuracy rates went up for the confidence interval, making it the best method of estimation amongst the Logistic Regressions.

As stated before, we were not able to create a Bag of Little Random Forests because of the time crunch, but based on our accuracy rates for the classification tree and the Random Forest, we would expect to get even better results with the Bag of Little Random Forests. When compared to the BLBGLM we get better results from Random Forest. This implies that there might be more power to the Random Forest approach because the accuracy is higher even though the sampling size was lower for Random Forest when compared to BLBGLM (B=100 vs B=1000). For the future, we now know that in the case of data with a lot of categorical variables and levels Bag of Little Random Forests would be an effective approach.

For now we can conclude that the best method to reduce our estimates, in terms of the Bag of Little Bootsraps with Logistic Regression, is the confidence interval approach (in the above table under "CI Rate (Upper)" and "CI Rate (Lower)"). So in the case of obtaining an estimate for our classification, we can use the following betas to get a lower bound and an upper bound estimate for the odds.

Income_LowerBound= -11.243 + age*.043 + raceAsian-Pac-Islander*1.064 + raceBlack*1.000 +raceOther*-1.189 + raceWhite*1.419 + sexMale*.338 + education_num*.338 +hours_per_week*.0298

Income_UpperBound= -10.282 + age*.0486 + raceAsian-Pac-Islander*1.981 + raceBlack*1.837 +raceOther*.319 + raceWhite*2.221 + sexMale*1.267 + education_num*.368 +hours_per_week*.0359

To get the probability simply use the following equation:

Prob_LowerBound = exp(*Income_LowerBound*)/(1+exp(*Income_LowerBound*)

Prob_UpperBound= exp(*Income_UpperBound*)/(1+exp(*Income_UpperBound*)

Finally, if the probability is higher than .5 the individual can be classified as the " >50K" income group, if it's lower than they can be classified as " <=50K".