Project Report

Ajinkya Rode ar2443@rit.edu
Harshit Shah hrs8207@rit.edu
Ninad Ingale nsi3177@rit.edu
Shikha Soni srs6573@rit.edu

Computer Science Department
Rochester Institute of Technology
Rochester, NY 14623

1 Abstract

In this paper we present analysis and predictions made using the census income dataset of a country. We have analyzed various hypothesis like the unemployment reasons, region wise industrial development, income tax to the country and few more that are important for making the economical decisions in country. We have used R and Weka for our analysis purpose and PSQL as a database management system.

2 Introduction

Many kinds of datasets are available online but we chose census dataset of a country, considering that there are many things that can be taken under analysis. Various patterns on the basis of their economic information can be predicted. The raw data set that was taken by us had around two hundred thousands entries of the country population including 45 columns like age, sex, education, tax information, and more information on the economic details.

On careful study of the datasets, we came up with hypotheses such as regionwise industrial and occupational patterns of the country, tax collection, finding out employment status. We came up with these hypotheses in the virtue of understanding various patterns in the populations economic status. We have built various graphical representations using R, while the classifications have been done in Weka. We have used the J48 algorithm for the classifications. It is a decision tree based algorithm, that produces univariate trees. Also, the J48 classifier displays additional information, including a text representation of the tree it uses to perform evaluations.

3 Design Considerations

3.1 Design Consideration for Database

1. **ER Diagram:**

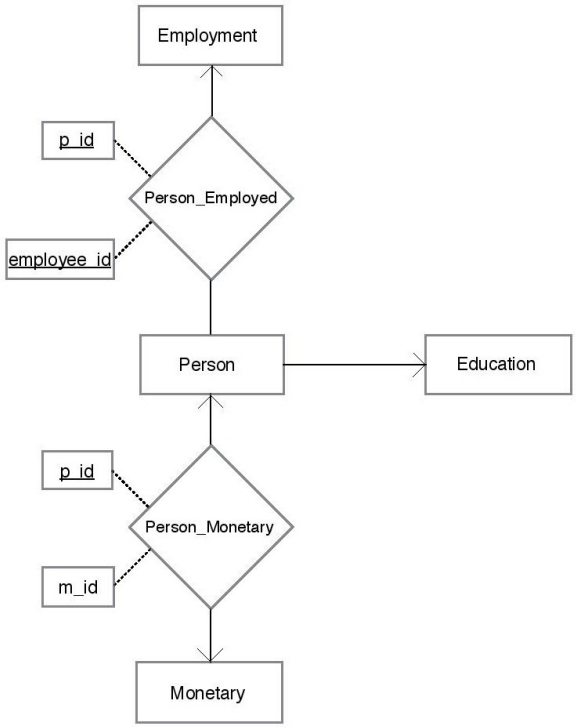

Figure 1: Census Income - ER Diagram

2. **Relations, Assumptions, Data Dictionary:**

   - **Relations:**

Person (p_id, age, education, marital_status, race, hispanic_origin, sex, labor_union, unemployment_reason, region_prev_residence, state_prev_residence, household_family_stat, live_in_house, citizenship, self_employed, veteran_benefits, cob_father, cob_mother, cob_self)

Employment (employee_id, class_worker, wage_phour, industry_code, occupation_code, employment_status, no_persons_worked, weeks_year)

Person_Employed (p_id, employee_id)

Education (education, edu_high_qual)

Monetary (m_id, capital_gain, capital_loss, dividend_stock, tax_status)

Person_Monetary (p_id,m_id)

- **Data Dictionary:**

  Here we will briefly discuss the important relations and define the ambiguous column names.

  → **Person:**

  We have created this relation to store basic personal information like age, education, sex, marital_status, etc...

  -*cob_father*: Fathers country of birth

  -*cob_mother*: Mothers country of birth

  -*cob_self*: Persons own country of birth

  -*household_family_stat*: Detailed household and family status of a person

  -*labor_union*: If a person is a member of a labor union. (Yes, no or not in universe)

  -*live_in_house*: If a person lives in the same house 1 year ago

  -*region_prev_residence*: Region of previous residence of a person (east, west, etc)

  -*state_prev_residence*: State of previous residence of a person (east, west, etc)

  -*unemployment_reason*: Reason for unemployment if person is not working

  -*veteran_benefits*: Person is entitled to veterans benefits or not.

  -*weeks_year*: Number of weeks worked in a year

  → **Employment:**

  This relation is created to store basic all employment details of a person. This table only contains information about persons who all are employed.

  -*class_worker*: Class of a worker like federal government, local government, etc

  -*employment_status*: Employment status full time, not in labor force, children and armed forces, etc...

  -*industry_code*: Major Industry Code

  -*no_persons_worked*: Number of persons working for an employer if it is an employer

  -*occupation_code*: Major Occupation Code

  → **Monetary:**

  This relation stores monetary information about an individual. For instance, capital gain/losses and tax filer details.

  -*capital_gain*: Capital gains of a person for the specific year

  -*capital_loss*: Capital losses of a person

  -*dividend_stock*: Details about dividends and stocks of a person

  -*tax_status*: tax filer status of a person

  → **Education:**

  Education relation contains person's highest education details and map the highest education with last attended institute. E.g. $9th$ standard → School.

  edu_high_qual - Highest education attended like school, college or University

  education - Highest qualification of a person

  → **Person_Employed** and **Person_Monetary** relations displays the relationship between *Person* and *Employed*, and *Person* and *Monetary*, tables respectively.

- **Assumptions:**

  → The attributes, employee_id and p_id are foreign keys in Person_Employed table.

→ Attributes m_id and p_id are foreign keys in Person_Monetary table.

→ In the Education table education is a primary key and it is a foreign key in Person table.

→ One person can be employed for more than one employer.

3. **Changes made to existing dataset:**

→ First we have removed columns which contain ambiguous values like *industry_recode*, *occupation_recode*. These columns contain numeric values which we were not able to figure out in regards to the industry recodes and they were thus not useful for our analysis.

→ Another columns *instance_weight* and *ignored* were removed since the former contained numeric values and later had flag values thus making them of no use. So, after removing these columns we were left with 36 attributes.

→ Furthermore there were columns like *mcc_msa*, *mcm_reg*, *msg_reg*. These columns too had values like '?' making them irrelevant. We removed them along with the other not useful entries. So, we removed these columns and left with 31 columns.

→ Finally, we have normalized our dataset into six above relations.

3.2 Design Consideration for Analysis

Our data set was of the CSV file format which was easy to load in both R and PSQL. Working on the hypothesis using R was a good option since it is a very good analysis tool which also has amenities to pass SQL queries and make changes in the data. We used the 'sqldf' package in R for SQL like querying. But there were certain problems that were faced in using R.

→ **Problems faced in R for SQL operations**:
- Performing DML operations like updating and inserting were difficult
- Applying integrity constraints is not possible
- Creating a database schema is not possible.

To overcome these we decided to use PSQL, for various table related problems, while continue to use R for analytical purposes, and graphical representations. PSQL which is more flexible and has proper data management capabilities was a good option. The whole data was loaded into PSQL after designing the schema and applying the integrity constraints.

→ **PSQL:**
- Inserted the whole data, applying various constraints.
- Removed tuples with the values containing 'Not in universe' Also values like NA were taken care of. Removed these entries since no useful data could be extracted from them.
- The total data dropped down to around 100,000 and this was divided into 6 different tables.
- The tables were studied and normalized into relations.
- We added a few columns and made changes o existing columns according to the hypothesis requirements. These will be stated along with each hypothesis .

We have used R to analyse the data from different views to identify certain patterns and Weka to obtain specific classification on these extracted $CSVs$.

In a brief we are using PostgreSQL for data extraction and cleaning, R for analysis and Weka for predictive analysis.

4 Analysis and Hypothesis

4.1 Employment Status:

In this analysis we have deduced if a particular person is employed or not based on his/her age, sex, education, citizenship, and marital status information.

→ **Approach**
- Columns used: *education*, *age*, *sex*, *citizenship*, *marital_status*, *employment_status*

- Modifications made to dataset

1) Grouped all *age* records into four categories, 18, 27, 58 & 65. Because below the age of 18, the population mainly consists of children, while everybody at the age of 65 or above are the retired ones. Other two groups are 27 and 58. These are made since the number of people with ages around 27 and 58 were in a majority.

2) Grouped *education* records into three categories. 'School Level','Bachelor's Degree', 'Master's-Doctorate Degree'.

3) Grouped *citizenship* records into two categories. 'US Citizen', 'Not a US Citizen'.

4) Grouped $marital_status$ records into two categories. 'Married', 'Unmarried'.

→ **Classification**

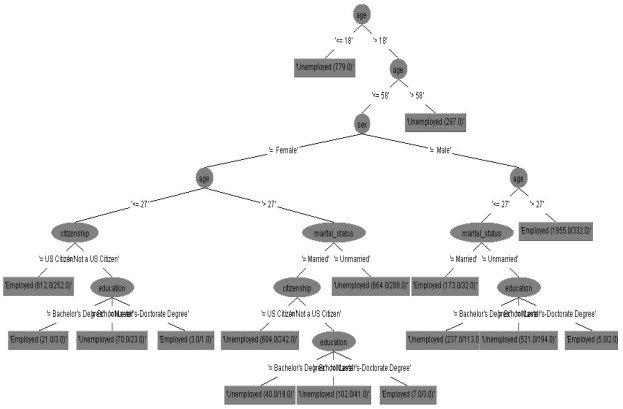

Figure 2: Employment Status

→ **Results**

It clearly predicted that the population of ages 18 were unemployed since they were children still pursuing education. While the ones above 58 were unemployed possibly since they had retired.

1) *Females:*

The model strangely predicted that females of the age greater than 27, and are unmarried will be unemployed, while the married US citizens are unemployed. The non US citizens are employed only with a Masters/doctorate degree, while others are unemployed.

Females of ages lesser than 27 and US citizens are employed, while the non US citizens with bachelors degree and Masters/doctorate degrees are employed while the ones with school degree are unemployed.

2) *Males:*

Males of age greater than 27 are employed irrespective of their educations. Males with age less than 27 and greater than 18, and if married then all are employed while only unmarried with Masters/Doctorate degree are employed. So they might still be pursuing their higher education.

The changes in the person's marital status, citizenship, and education majorly affects the person's employment status.

4.2 Preferred gender in a given industry:

This hypothesis predicts the preferred gender for any industry on the basis of their age and education. Each industry has their preferences of sex when it comes to the education level, while there are certain industries which prefer either of the sexes, irrespective of the two factors.

→ **Approach**

- Columns used: *education*, *age*, *sex*, *education*, *industry_code*

- Modifications made to dataset

1) The education levels and ages are grouped on the same lines as in the above hypothesis.

2) The industries count was plotted and they were divided into 5 groups on the basis of the descending order of the counts. This graph was plotted in R and shown as below.

→ **Graph:**

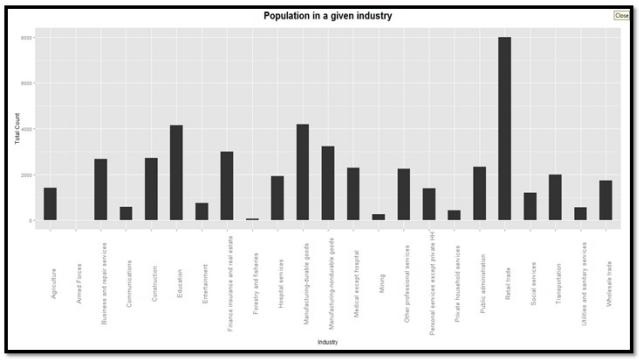

Figure 3: Population in the given Industry

3) From the above graph, we grouped all the industries into 5 groups and performed our classification as below.

Industries Group 1:

Industry	Count
Manufacturing-durable goods	9015
Retail trade	17070
Education	8283
Manufacturing-nondurable goods	6897

Industries Group 2:

Industry	Count
Finance insurance and real estate	6145
Construction	5984
Business and repair services	5651
Medical except hospital	4683
Public administration	4610

Industries Group 3:

Industry	Count
Other professional services	4482
Transportation	4209
Hospital services	3964
Wholesale trade	3596
Agriculture	3023

Industries Group 4:

Industry	Count
Personal services except private HH	2937
Social services	2549
Entertainment	1651
Communications	1181
Utilities and sanitary services	1178

Industries Group 5:

Industry	Count
Private household services	945
Mining	563
Forestry and fisheries	187
Armed Forces	36

→ **Classification**

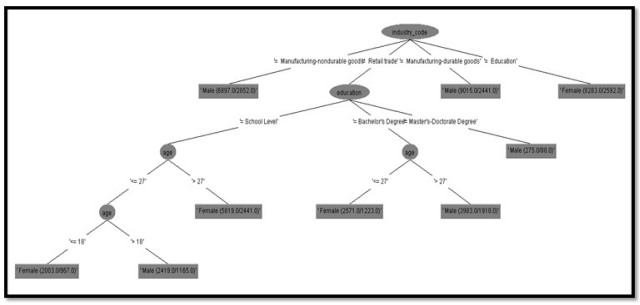

Figure 4: Industries Group 3

→ **Results**

Looking at the tree in Figure 2, it appears that

Manufacturing non Durable/ Durable goods: Prefers only Males in industry.

Retail trade: Females with school level education and Bachelor level are preferred, provided their ages are less than 18 and greater than 27. Males with ages greater than 18 but less than 27 having school level education, are preferred. Males with the bachelors degree with the age of greater than 27, or with masters degree are preferred.

Education: This industry only prefers the female population.

The accuracy of our model is 61.9847 %, which is achieved by training our model on 41265 instances. This accuracy is not up to the mark because at the very beginning data was cleaned for 'Not in universe' entries and that was a huge amount of data.

In similar way, other four classifications are performed on all other industries groups. Due to space and image quality constraints, we have not included other four decision trees of same hypothesis. From them we can predict the same results as above group, which industries prefer males or females.

4.3 Region wise Industrial Development:

In this hypothesis, region wise industrial development of a country is analyzed based on industry_code in all four regions. The most influential industry in the country is selected and based upon the occupation_code for that particular industry_code, further analysis is done i.e the majority of population in a particular occupation in a particular region is found.

→ **Approach**

- Columns used: *industry_code*, *curr_region*, *state_curr_residence*, *occupation_code*

- Modifications made to dataset:

1) A new column, 'state_curr_residence' is added with 48 states randomly assigned to employee relation.

2) A new column, 'curr_region' is added containing four regions; North, South, East and West generated on the basis of 'state_curr_residence'.

3) These added columns, *state_curr_residence* and *curr_region* to support our hypothesis and make our analysis more precise.

→ **Graph:**

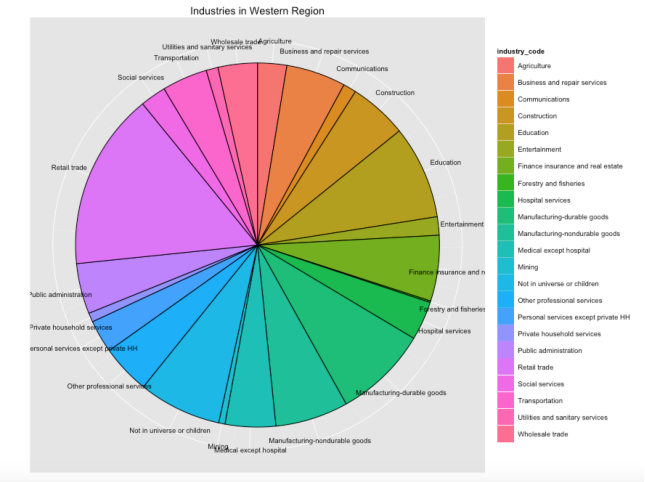

Figure 5: Major industry in a Region

→ **Graph:**

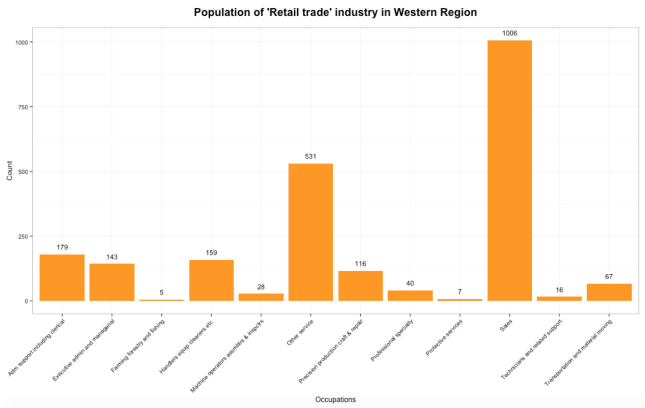

Figure 6: Major occupation in a Region for influential industry

→ **Results**

1) From figure 5 we conclude that the major industry in the country in the Western Region is 'Retail Sales'

2) Similarly we generate most influential industry in all four regions. We predict from the figure that 'Retail Sales' is the most influential industry in the whole country.

3) To make this more accurate we drilled down further to find out which all occupations in this industry are mostly preferred

4) From further analysis we found that sales is the major occupation in the Western Reagion. This pattern is followed for all four regions.

5) Hence to conclude, 'Retail Sales' is the major flourishing industry and most of the population chose sales as their occupation in the country.

4.4 Tax Analysis in a country:

In this analysis we calculate gross income of the country using the wage_phour and further calculate the income tax of an individual based on his tax_filer_status and self generated income-tax rules.

→ **Approach**

- Columns used: $wage\_phour$, tax_filer_status, $weeks\_year$

- Modifications made to dataset:

1) A new column, 'annual_income_in_terms_of_100k' is added and the annual income of an individual is calculated based on the formula:
$annual\_income\_in\_terms\_of\_100k = (wage\_phour * 40 * weeks\_year)/100000$

2) Here we assume that an individual works for 40 hours per week.

3) The tax that an individual is levied is calculated based on some general tax rules[2].

→ **Graph:**

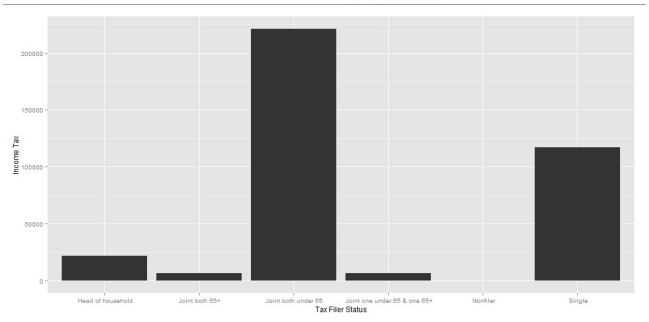

Figure 7: Income Tax of the Country

→ **Results**

1) The most tax paying tax_filer_group is 'Joint both under 65'. This is because there is a majority of people from group Joint both under 65 in the dataset at hand.

4.5 Industry wise Average Annual Income:

In this analysis we find out the average annual income of a male and female and based on that we find

a pattern with respect to industry_code in the country.

→ **Approach**

- Columns used: *wage_phour*, *sex*, *weeks_year*, *industry_code*

→ **Graph:**

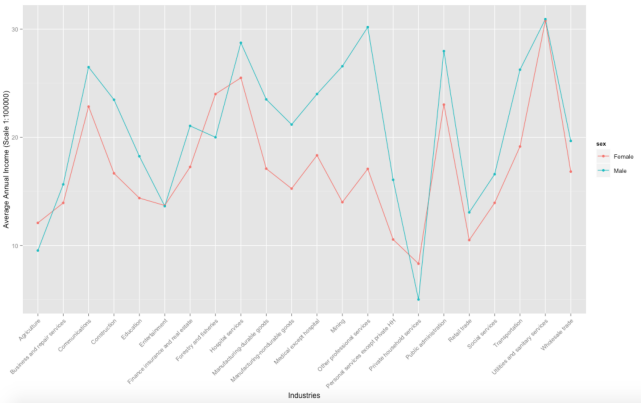

Figure 8: Industry wise Average Annual Income

→ **Results**

1) Except for fields like Agriculture, Forestry, and Private household services all the industries have males with higher annual income.

2) The industry Utilities and Sanitary services does pay its employees better than other industries.

3) The reasons for females earning higher in these industries are unpredictable.

4) Almost all industries have females with less annual income in each industry by a large margin.

4.6 Age wise education for particular gender:

The above graph has been plotted to compare the highest level of education of the population, dividing them sex wise (Male, Female) and age wise(18,27,58,65). Here we are analyzing age wise educated people and comparision between males and females.

→ **Approach**

- Columns used: *sex*, *education*, *age*

- Modifications made to dataset

1) Grouped all *age* records into four categories, 18, 27, 58 & 65 as above hypotheses.

2) Grouped *education* records into seven categories from total availabe sixteen categories for better analysis. Because it is more accurate analysis if we have less categories. The new seven categories are 'Pre School-Kindergarten','Elementary School (1st-4th grade)', 'Middle School (5th-8th grade)', 'High School (9th-12th grade)', 'College or University (Bachelor's Degree)', 'College or University (Master"s-Doctorate Degree)', 'Vocational-Associates Degree'.

→ **Graph**

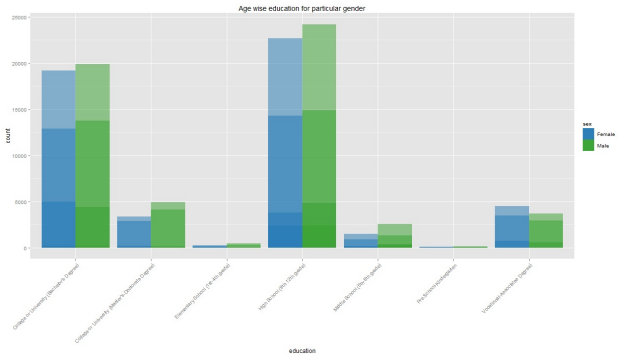

Figure 9: Age wise education for particular gender

In figure 9, the stacked color variations, represents the age group. Where the darkest, at the bottom represents age group of 18, while the lightest, at the top represents age group of 65.

→ **Results**

1) On studying the graph one obvious thing was covered that the population of ages less than 18 are limited till high school and college university.

2) On observation the education levels in each group of both the sexes were almost similar. Majority of the population studied till either high school or completed their bachelor's at the most.

3) People of the age group 58 have studied till their high school or have attained bachelors degree.

5 Future Work and Lessons Learned

Working on this project made us familiar with how exactly should a dataset be approached. The various data mining and analysis techniques. We understood how exactly should the bad data be handled, like cleaning the data and getting over the glitches in the important parameters. Exploring tools like R and Weka were extremely helpful. The major challenge we faced was bad data entries. That was overcome and the data was grouped in sets to plot the trees

better.

In future we can train our model better by improving the accuracy. This can be done if we group the data better, also if we train the model better. Making a user interface for the employment status, unemployment reasons and possible placement of industries on the basis of age and education can be designed to display the project in an interactive way.

6 Conclusion

The project is used to analyze various aspects of the country's economic sector. In conclusion we found out that the employment status of the person, the tax status, region wise industries, preferred sex by a given industry, sex wise annual income. The hypothesis that states the tax information helps find the sector that gives maximum tax thus helping the country predict the tax revenues that could be expected in future. Another important information was where the development of each industry and occupation in the 4 regions could be found, thus helping the nation predict the flourishing states, industries as well as occupations in the country.

References

[1] Lichman, M. *UCI Machine Learning Repository.* Morgan Kaufmann, (2013). `https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)`

[2] *National Tax Services, Inc.* 2005. `http://www.unclefed.com/IRS-Forms/1995-1991/1994TaxRateSchedules.pdf`