FINAL PROJECT REPORT

Income prediction using 1994 and 1995 US census data

Submitted By:

Student Name	SIS ID
Ketan Chaudhari	56SL2MZSE9
Mayur Khandeshe	XJE1N4EQ2A
Mohd Zuhaib	0E573X5AIO
Rahul Chavan	BZRD8TQLU1
Shyam Sabbi	RAVN6V4SCN

Batch: DSE_PUNE_SEPT2019

Mentor: Ankush Bansal

GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI

greatlearning
Learning for Life

Abstract

A country is made up of every individual in a society and status of the society depends upon the income of an individual. Income affects every aspects of society like Health, life expectancy, social stability, crime rate, population-wide satisfaction and happiness. Thus, income of an individual plays a key role in the development and prosperity of the whole country. But income of an individual depends upon lots variables, just to mention a few Education and Occupation, Preowned business, Area of living, migration and government benefits. All such information is collected from census. A census is the procedure of systematically acquiring and recording information about the members of a given population. This term is used mostly in connection with national population and housing censuses; other common censuses include traditional culture, business, supplies, agricultural, and traffic censuses. The United Nations defines the essential features of population and housing censuses as "individual enumeration, universality within a defined territory, simultaneity and defined periodicity".

For this project, we have been provided with the data set which contains weighted census data extracted from the 1994 and 1995 ‘Current Population Surveys’ conducted by the U.S. Census Bureau. It contains data about all the civilian non-institutional population of the United States living in housing units and members of the Armed Forces living in civilian housing units on a military base or in a household not on a military base. Prediction task is to determine the income level for the person represented by the record. Incomes have been binned at the $50K level to present a binary classification problem. There are several methods for binary classification like Logistic Random Forest, Decision Tree, SVC and boosting. We are going to apply every model possible to get better results and also for this interim report we have included base model results which we found out using random forest.

Acknowledgements

At the outset, we are indebted to our Mentor Mr. Ankush Bansal for his time, valuable inputs and guidance. His experience, support and structured thought process guided us to be on the right track towards completion of this project.

We are extremely gifted and fortunate to have Ms. Varsha Mali as our Technical Guide. Her in-depth knowledge coupled with her passion in delivering the subjects in a lucid manner has helped us a lot. We are thankful to her for her guidance towards entire coursework.

We also thank all the course faculty of the DSE program for providing us a strong foundation in various concepts of analytics & machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving us the necessary support, space and time to complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Ketan Chaudhary
Mayur Khandeshe
Mohd Zuhaib
Rahul Chavan
Shyam Sabbi

Date:26th Feb 2020

Place: Pune

Certification of Completion

I hereby certify that the project titled “Income prediction using 1994 and 1995 US census data” was undertaken and completed under my guidance and supervision by Ketan Chaudhary, Mayur Khandeshe, Mohd Zuhaib, Rahul Chavan and Shyam Sabbi students of the Sept 2019 batch of the Post Graduate Program in Data Science & Engineering, Pune.

Mr. Ankush Bansal

Date: 26th Feb 2020

Table of Contents

Table of Figures

Abbreviations

Abbreviation	Expansion
LR	Logistic Regression
ROC	Receiver Operator Characteristic
LGBM	Light Gradient boosting method
XGB	Extreme Gradient Boosting
GB	Gradient Boosting
KNN	K Nearest Neighbor
ADB	Ada Boost
mRMR	Minimum Redundancy and maximum relevance
SMOTE	Synthetic Minority Oversampling Technique
TPR	True Positive Rate
FPR	False Positive rate

Executive Summary

Background: The dataset is a part of the census of United States held in 1994 and 1995. The record is very huge and the dataset accounts for approximately 3 Lakhs data entry. The data is varied and represents different sections of the society. The data contains both numerical and categorical data. The data is not cleaned and contains noisy elements. Data needs to be pre-processed according to the objective of the project.

Objective: The main objective of the project is to classify people earning <=50k or >50k annually based on several explanatory factors affecting the income of an individual like Age, Occupation, Education, etc.

Scope: The census tells us who we are and where we are going as a nation, and helps our communities determine where to build everything from schools to supermarkets, and from homes to hospitals. It helps the government decide how to distribute funds and assistance to states and localities. Many businesses would like to personalize their offer based on customer's **income**. High-**income** customers could be, for instance, exposed to premium products. As a customer's **income** is not always explicitly known, predictive model could estimate **income** of a person based on other information.

Approach & methodology: The data given is highly noisy and imbalanced. Preprocessing steps include treating null values according to the context. Removing outliers or capping them to avoid data loss. Transforming continuous variable for outlier treatment. Feature Engineering to combine various columns providing similar information. Bucketing different categories resembling similar properties is also a part of feature engineering. Performing exploratory data analysis and statistical significance hypothesis testing with respect to target variable to understand the factors effecting the target variable most. EDA also helps in checking variance of the data and helps expose intra-correlated features which can be dropped before final model building.

Introduction

The census provides information on size, distribution and socio-economic, demographic and other characteristics of the country's population. The data collected through the census are used for administration, planning and policy making as well as management and evaluation of various programmes by the government, NGOs, researchers, commercial and private enterprises, etc. Census data is also used for demarcation of constituencies and allocation of representation to parliament, State legislative Assemblies and the local bodies. Researchers and demographers use census data to analyze growth and trends of population and make projections. The census data is also important for business houses and industries for strengthening and planning their business for penetration into areas, which had hitherto remained, uncovered.

To collect such an important data of individuals, US government has established US Census Bureau in 1902. The Census Bureau is the federal government's largest statistical agency. They are dedicated to providing current facts and figures about America's people, places and economy. To get such facts and collect data US government conducts surveys. They also collect additional data from other sources. Primary sources for additional data are federal, state, and local governments, as well as some commercial entities. Following are surveys conducted under the administration of US Census Bureau:

- The **American Community Survey (ACS)** is an ongoing annual survey that shows what the U.S. population looks like and how it lives. The ACS helps communities decide where to target services and resources.
- **Demographic surveys** measure income, poverty, education, health insurance coverage, housing quality, crime victimization, computer usage, and many other subjects.
- **Economic surveys** are conducted monthly, quarterly, and yearly. They cover selected sectors of the nation's economy and supplement the Economic Census with more-frequent information about the dynamic economy. These surveys yield more than 400 annual economic reports, including principal economic indicators.
- **Sponsored surveys** are demographic and economic surveys that they conduct for other government agencies. This include the Current Population Survey, the National Health Interview Survey, and the National Survey of College Graduates.

We have been provided with such a Census data from 1994-95 American Community Survey (ACS), Demographic survey and Economic survey. The universe in this data is the civilian non-institutional population of the United States living in housing units and members of the Armed Forces living in civilian housing units on a military base or in a household not on a military base. This data provides the usual monthly labor force information, but in addition, provides

supplemental data on work experience, income, noncash benefits, and migration. Comprehensive work experience information is given on the employment status, occupation, and industry of persons 15 years old and over. Additional data for persons 15 years old and older are available concerning weeks worked and hours per week worked, reason not working full time, total income and income components, and residence on March 1, 1995. Data on employment and income refer to the preceding year, although demographic data refer to the time of the survey. Characteristics such as age, sex, race, household relationship, and Hispanic origin are shown for each person in the household enumerated.

Problem Statement and Project Scope:

Prediction task is to determine the yearly income level for the person represented by the record in the US Census data from 1994-95. Incomes have been binned at the $50,000 level to present a binary classification problem. We have to predict whether a person represented by a record in a dataset have yearly income 'above $50,000' or 'below $50,000'.

Data Source:

Current Population Survey, March 1994 & March 1995 Technical Documentation. The documentation includes this abstract, pertinent information about the file, a glossary, code lists, and a data dictionary. All the census data is available from Administrative and Customer Services Division, Customer Services, Bureau of the Census, Washington, DC 20233. Available on https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar95.pdf.

Data preparation:

The source dataset received has been prepared to ensure that the fields are cleaned up, the values are suitable for model building and the variable names are self-explanatory. The broad approach for data preparation can be outlined as:

- Data pre-processing steps: Label Encoding, Outlier Treatment, Standardization, Oversampling
- Box plot is drawn for Independent features against Target variable and outlier had been detected.
- Standard Scalar function from Scikit learn library since the numerical variable are of different scale in order to obtain better performance.
- Since our Dataset is highly imbalanced, we used SMOTE oversampling technique in order to tackle class imbalance. Weekend and Revenue feature is converted into binary value 0's and 1's Since the outliers are legitimate, we have decided to retain them in data

Statistical tools & techniques

Various classification algorithms have been used to for the classification problem.

The model building exercise has also considered cross validation and tuning techniques to ensure that the models built perform well when used for prediction.

The classification algorithms used for the prediction:

● Logistic regression ● Ada Boost ● Random Forest ● Light Gradient Boosting

Model performance measures used for evaluating models:

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. The choice of the right model performance measures is highly critical since the dataset is a highly imbalanced dataset and the minority class is 6%. Model accuracy alone may not be enough to evaluate a model. Hence the following model performance measures have been used to evaluate the models, based on the confusion matrix built for the predictions on the training and test datasets:

Table 1 : Confusion Matrix Format

	Negative (Predicted)	Positive (Predicted)
Negative (Observed)	True Negative (TN)	False positive (FP)
Positive (Observed)	False negative (FN)	True positive (TP)

Accuracy : It is the number of correct predictions made by the model by the total number of records. The best accuracy is 100% indicating that all the predictions are correct. Accuracy is not a valid measure of model performance. Even if all the records are predicted as 0, the model will still have an accuracy of 94%. Hence other model performance measures need to be evaluated. Sensitivity or recall

Sensitivity (Recall or True Positive Rate) : It is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall or true positive rate (TPR).

For our dataset, it gives the ratio of actual citizens who are actually earning above 50k and who are classified correctly.

Specificity (True Negative Rate) : It is calculated as the number of correct negative predictions divided by the total number of negatives. For our dataset, specificity gives the ratio of actual citizens earning below 50k by the number of citizens who are predicted to be earning less than 50k.

Precision (Positive Predictive Value) : It is calculated as the number of correct positive predictions divided by the total number of positive predictions. Precision tells us, what proportion of customers who generated revenue as customers actually generated revenue. If precision is low, it implies that the model has lot of false positives.

F1-Score : F1 is an overall measure of a model's accuracy that combines precision and recall A good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

ROC chart & Area under the curve (AUC) : ROC chart is a plot of 1-specificity in the X axis and sensitivity in the Y axis. Area under the ROC curve is a measure of model performance. The AUC of a random classifier is 50% and that of a perfect classifier is 100%. For practical situations, an AUC of over 70% is desirable.

Level of significance : For all the hypothesis tests in the project, the level of significance is assumed as 5% unless specified otherwise.

Data Preprocessing

Duplicated Values: Dropping the duplicated columns.

Category Merging: There are many instances where high number of categories are present in some of the features. Combining similar categories (according to the target variable) will reduce categories and will help in better model building.

1. Industry Code: Categorical data with 51 categories converted to 13 categories having similar industry niche. Instances with no industry code are separated in armed forces and children categories.
2. Occupation Code: Categorical data with 46 categories converted to 9 ordinal categories having similar level of position in the corporate earning ladder. Instances with no occupation code are separated in armed forces, children and no data categories.
3. Weeks Worked in Year: Discrete variable converted into 3 categories namely, non-working, part-time, and full-time. EDA of the new feature portrays clear variance with respect to target variable.
4. Age: Discrete variable converted to 6 categories with predefined age brackets affecting a salary of a person mentioned in the documentation of census.
5. Education: Reduced down to 6 categories like children, high-school, under-graduate, post-graduate etc.
6. Marital Status: Combined various categories to form 2 broad categories namely married and single.
7. Race: Combined minority races to form a minority race category.

Feature Engineering: Feature engineering is performed to combine features which are giving similar type of data. Combining features will naturally decrease noise and will create a better variance feature that'll perform better in prediction of target variable.

1. Combining two variables, capital gains and capital losses to form one feature which will be better correlated with the target variable.
2. Combining 4 variables giving information of mother's birthplace, father's birthplace, self-birthplace and citizenship. First step was to divide the data into two categories: native born and foreign born. Next step was to check each column and assign labels accordingly to the new feature. For example, If the whole family is American born, they're classified as native If both parents are foreigners it is classified as foreigner_parents_nativeborn. Similarly, other categories are created.

3. Dropping household stats feature because of a similar feature providing similar data with broader classification, hence less categories.
4. Adding a new feature employment status instead of reason for unemployment because of high null values. Created using features like: reason for unemployment, owns business or self-employed, major occupation code, age. By doing this we got most of the data which is missing in single column.
5. Deleting the columns with high percentage **null value**s > 75% and which can't be combined with any other feature to generate information.

Exploratory Data Analysis & Feature Engineering

Performing EDA on the various variables and with respect to target variables to get actionable insights.

Target variable:

The pie chart shows that only 6.35% of data is present in bin 'above 50k' and 93.65% of data is present in bin 'below 50k'. This means there is a huge class imbalance for the target column. We need to tackle this problem with Sampling the data, which we will do later on. The figure 1 shows the distribution of Target classes in the Pi-chart.

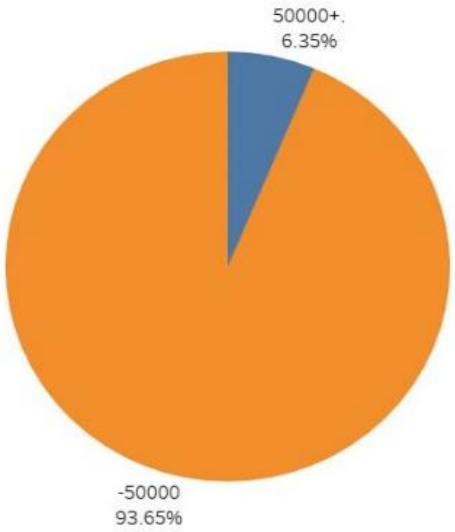


Figure 1 : Pi-chart Distribution of Target

AGE:

Age is distributed from 0 till 92years. We have bucketed the age data as follows:

1 : Below 15 years old, Children data
2 to 5 : 16 to 65 years old, Working Class
6 : Above 65 years old, Retired Class

The Figure 2 shows the Stacked Bar plot representation of Age (Binned).

Orange: Above $50,000
Blue: Below $50,000

Low age categories have negligible percentage of above 50k earners. Most above 50k earners in 4th category which is middle aged man.

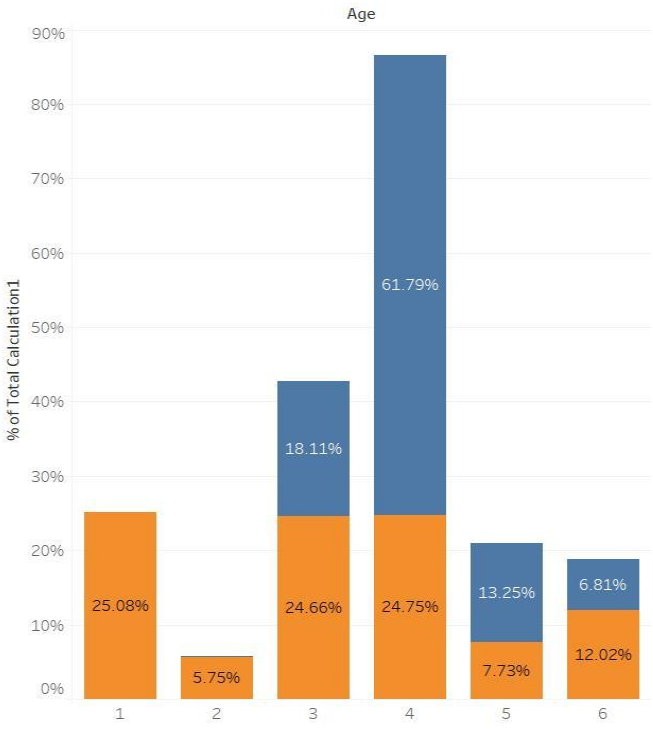


Figure 2 : Stacked Bar Plot representation of Age w.r.t. Target

Industry code (before label merging):

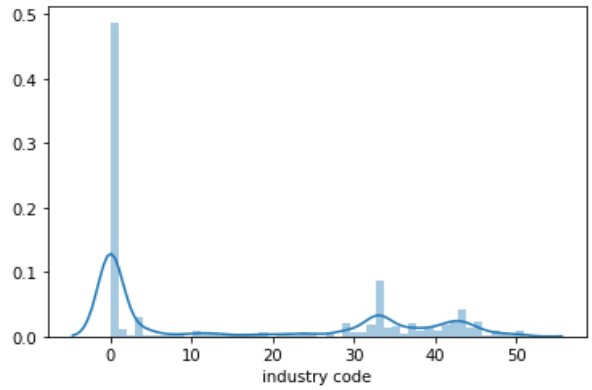


Figure 3 : Distribution Plot of Industry Code (Before Bucketing)

Industry code represents the code of the industry in which an employed individual works. According to the documentation, there are 236 categories for the employed, with 1 additional category for the experienced unemployed. These categories are aggregated into 51 detailed groups. This is categorical variable. Figure 3 shows the distribution of industry code. The distribution plot shows that most of data is present around group number 0.

Occupation code (before label merging):

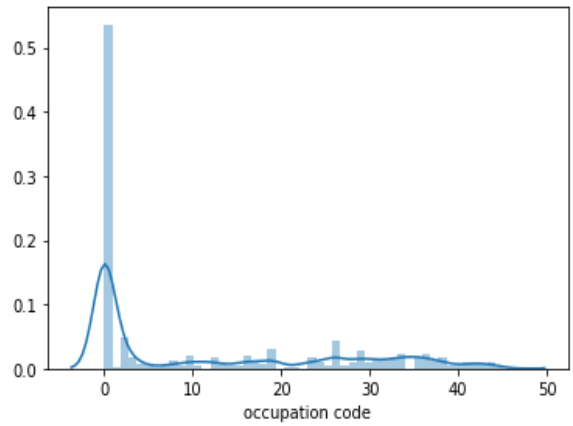


Figure 4 : Distribution Plot of Occupation Code (Before Bucketing)

Occupation code represents the code of the post / occupation of an employed / self-employed individual. According to the documentation, there are 500 categories for the employed with 1 additional category for the experienced unemployed. These categories are aggregated into 46 detailed groups. This is categorical variable. Figure 4 shows the distribution of occupation code. The distribution shows that Majority of data is present in group 0.

Wage per hour:

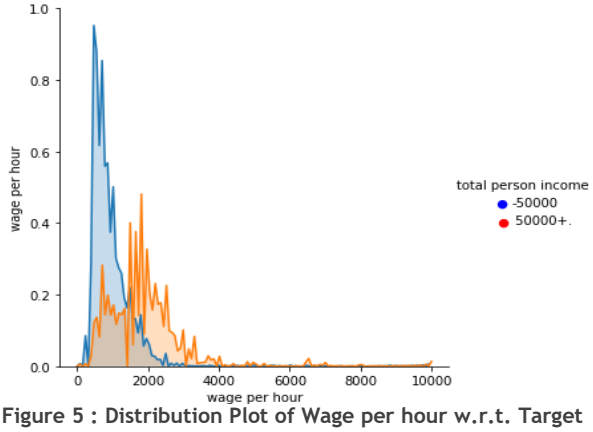


Figure 5 : Distribution Plot of Wage per hour w.r.t. Target

This shows hourly earnings of an individual if he is working. There are 2 major types of employments, Part time and full-time workers. Out of them, wage per hour represents usual hourly earnings from both the current job (Part Time) and earnings from the longest job (Full Time). Figure 5 shows the distribution of wage per hour. The distribution shows that Wage per

hour data is continuous and there is a clear overlap between both the categories of below and above 50k person income's segments. Wage per hour data is very much important to find out the income of an individual but distribution is very much right skewed, and we can use such data directly.

Major industry code:

Figure 6 shows the count plot w.r.t income. The count plot says that there are more than 5 industries which contribute more than 50% in deciding the >50K salary band. And *'Not in Universe and Children'* is the major class decider in <50K salary band.

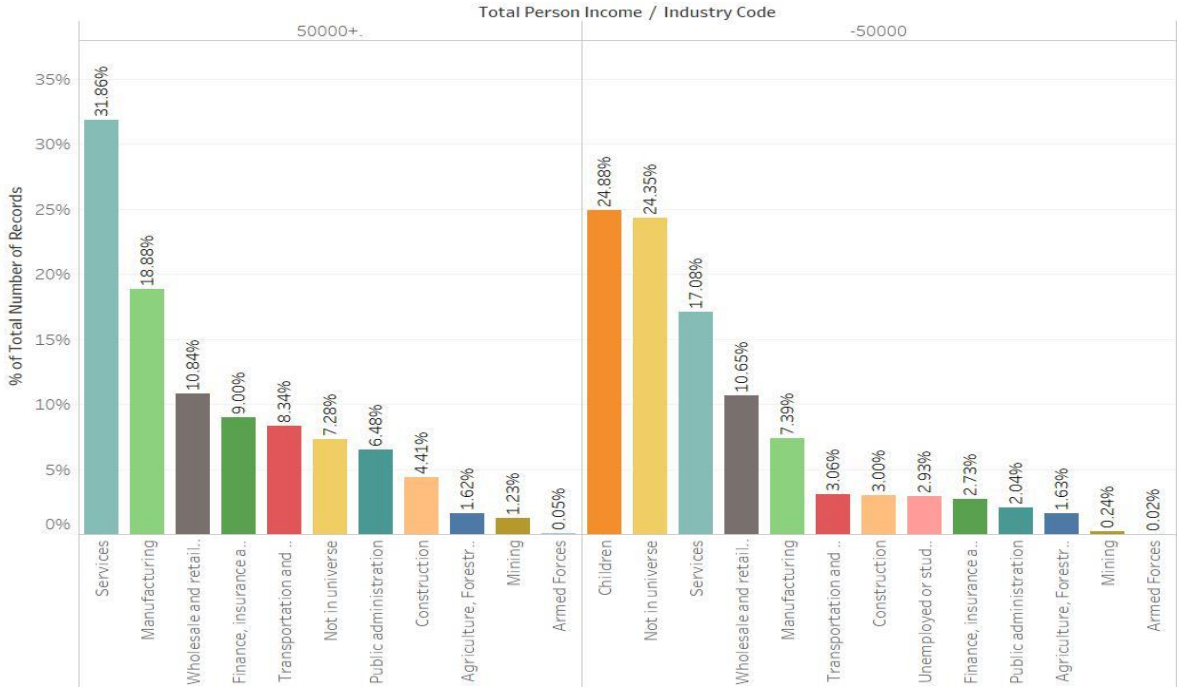


Figure 6 : Bar-plot Represetation of Major Industry Code (After Bucketing)

Major Occupation Code

The below Figure 7 shows the count plot of major occupation code with respect to income. From plot we can say that *'Executive admin and managerial', 'Professional specialty', 'Sales'* are the major class differentiators for above $50,000 band and for below $50,000 band again *'Not in Universe'* is major class. This is because *'Executive admin and managerial'* & *'Professional specialty'* is the highest paid occupation in the country. Whereas, *'Not in Universe'* is the class of data with no data available.

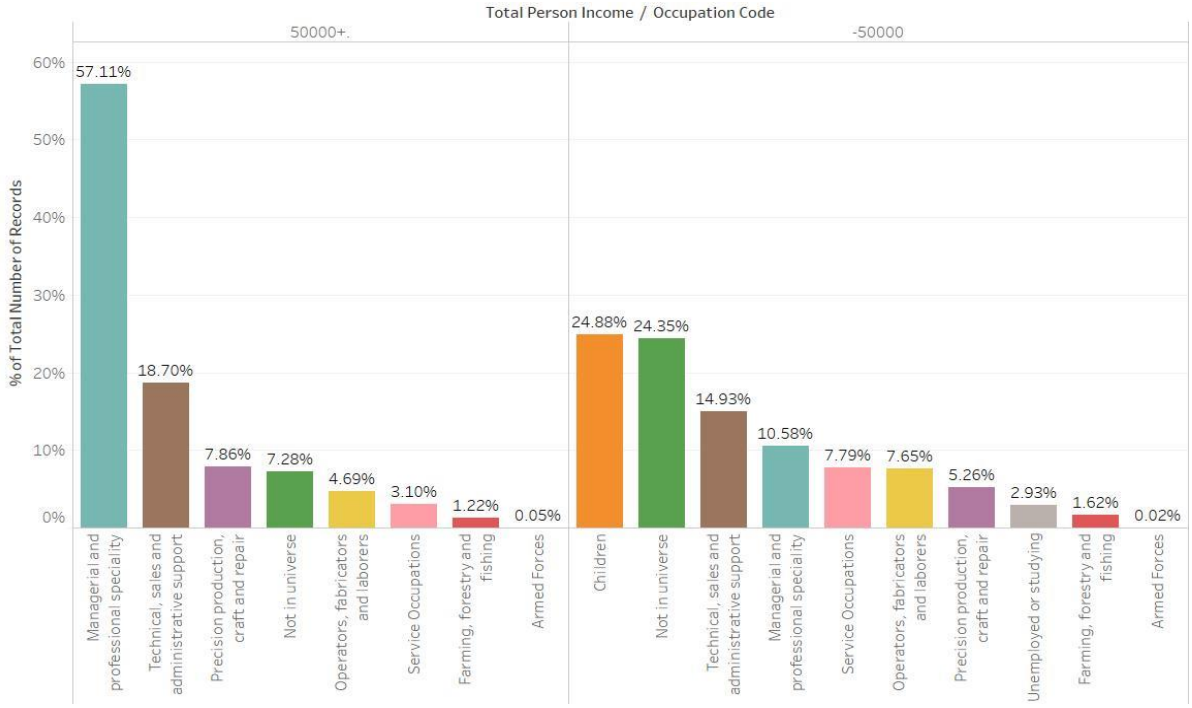


Figure 7 : Bar plot representation of Occupation Code (After Bucketing)

Class of worker

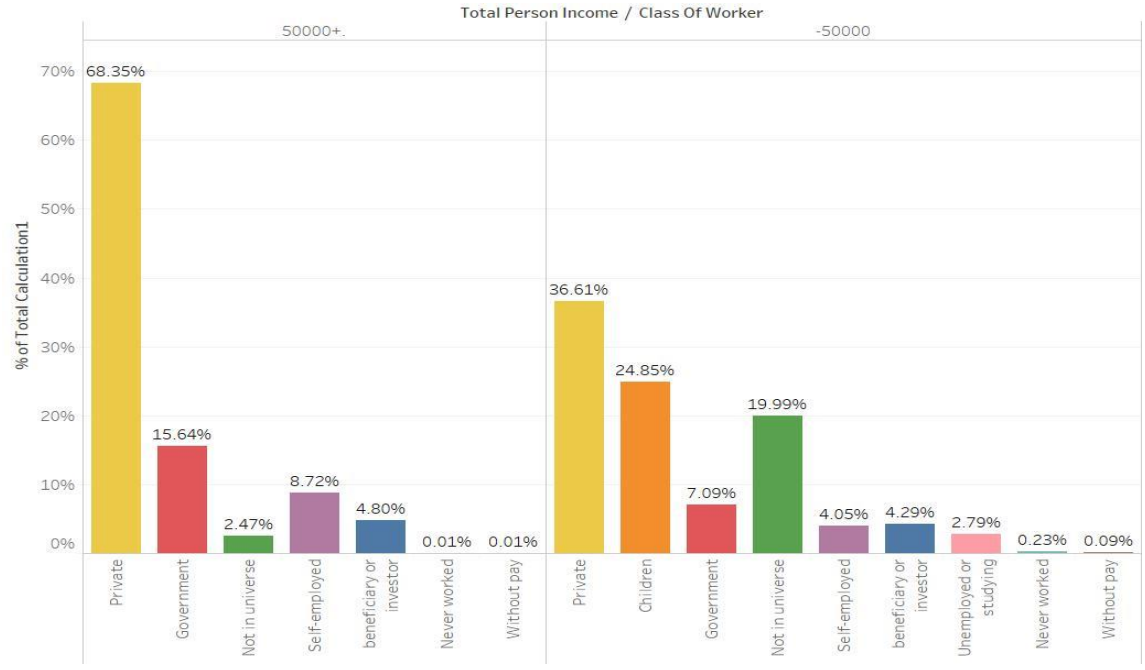


Figure 8 : Bar-plot Representation of Class of Worker (After Bucketing)

Above Figure 8 shows the bar-plot representation of Class of Worker with respect to Target Variable. It shows that Private more percentage in higher earning category versus lower earning category.

Member of a labor union

Fig 7 shows the count plot. From count plot we can say that Here no columns are showing any classification related data therefore after doing feature engineering we can conclude to keep or remove this feature.

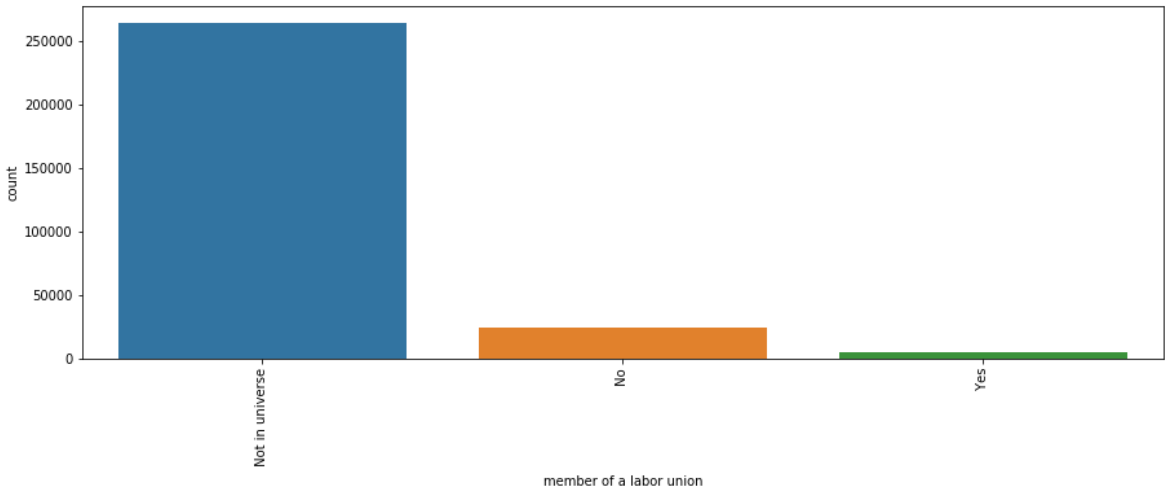


Figure 9 : Countplot of Member of a labor Union

Region of previous residence

This is a migration related data, it shows region before migration where was an individual living. Let's check the data first.

```
 Not in universe    92.012989
 South               2.489147
 West                2.084088
 Midwest             1.791147
 Northeast           1.372415
 Abroad              0.250214
Name: region of previous residence,
```

Figure 10 : Screen Capture of Variable percentage

92% of data is missing. So, it's after feature engineering we will check whether to keep or remove the feature.

State of previous residence

This is a migration related data, it shows state before migration where was an individual living. Let's check the data first.

```
 Not in universe    92.012989
 California          0.884293
 Utah                0.544180
 Florida             0.442318
 North Carolina      0.412237
Name: state of previous residence,
```

Figure 11 : Screen Capture of Variable percentage

92% of data is missing. So, it's after feature engineering we will check whether to keep or remove the feature.

Education:

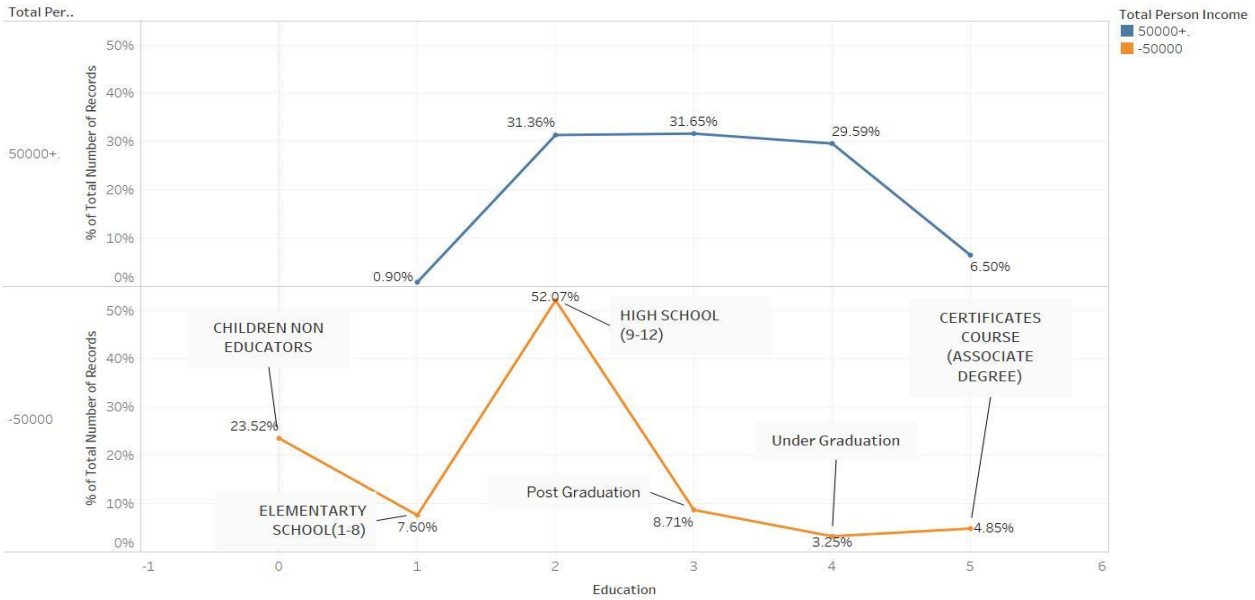


Figure 12 : Trend of Education w.r.t. Target

From the graph it can be inferred that low income group has high percentages of high school and non-educated students whereas the curve is evenly distributed for high level income group.

Full or part time employment stat:

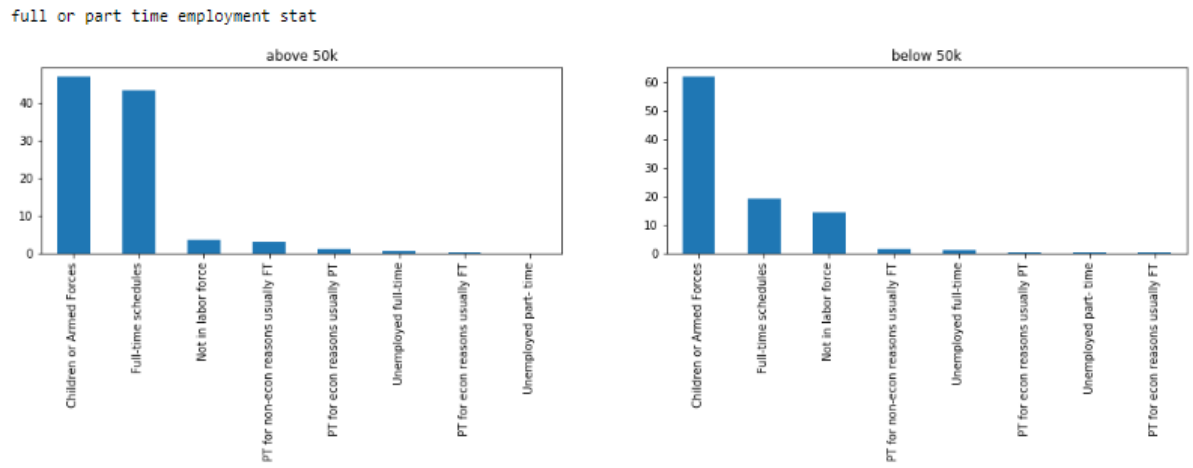


Figure 13 : Employment Status w.r.t. Target (After Bucketing)

Figure 13, from the above chart we can say that in **Above 50K,** children or armed forces earns more income , we are not having any relevant feature so that we can conclude how they are earning income, people who are working for full time schedule earns are almost 40%.

Whereas in **Below 50K,** children or armed forces earns more income , we are not having any relevant feature so that we can conclude how they are earning income.

Tax filer status:

Figure 14, From the below chart we can say that in **Above 50K,** 71% people filed tax return for joint both under 65 (A married couple can file a joint return or separate returns, You're legally married or, You're not legally separated under a divorce or separate maintenance decree.) followed by 18% people filed tax return for single (singles are :You're unmarried or legally separated from your spouse under a divorce or separate maintenance decree. You don't qualify to file as head of household or qualifying widow(er).) 3% peoples are filed tax return for head of household and joint both 65+ , in **Below 50K,** almost 40% people are non-filers,31% of people files tax return for joint both under 65 ,almost 20% people files for singles,4% people for joint both 65+,3% people for head of household

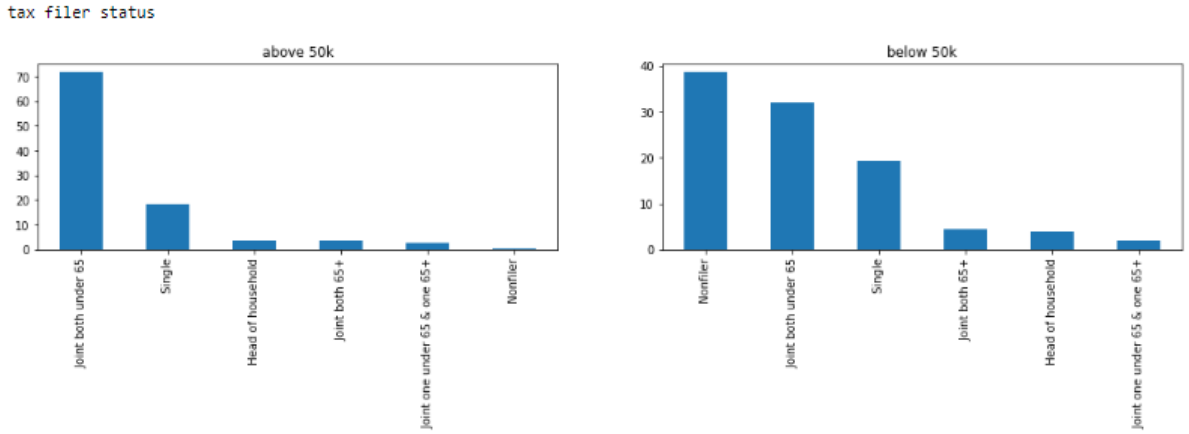


Figure 14 : Count-plot of Tax Filer Status

Veteran Benefits

Figure 15, From the below chart we can say that in **Above 50K,** almost 99% people have 2 (no) veteran benefits gets above 50K followed by 1% people have 1(yes) veteran benefits, no 0 veteran benefits. In **Below 50K,** almost 73% people are having 2(no) veteran benefits gets below 50k followed by almost 25% people are not having veteran benefits but earns below 50K, 1% people are having 1 (yes) veteran benefits.

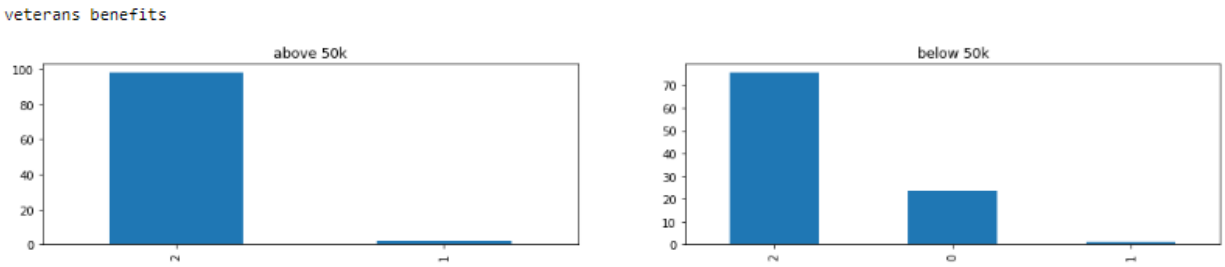


Figure 15 : Countplot of Veteran Benefits w.r.t. Target

own business or self employed

Figure 16. From the below chart we can say that in Above50K, almost 80% people data is not applicable (0),10% people data is not owns business, and 4% people data are having own business, whereas in below 50K, almost 90% people data is not applicable, 8% people data is not owns business, 2% people data owns business. **Very low importance feature.**

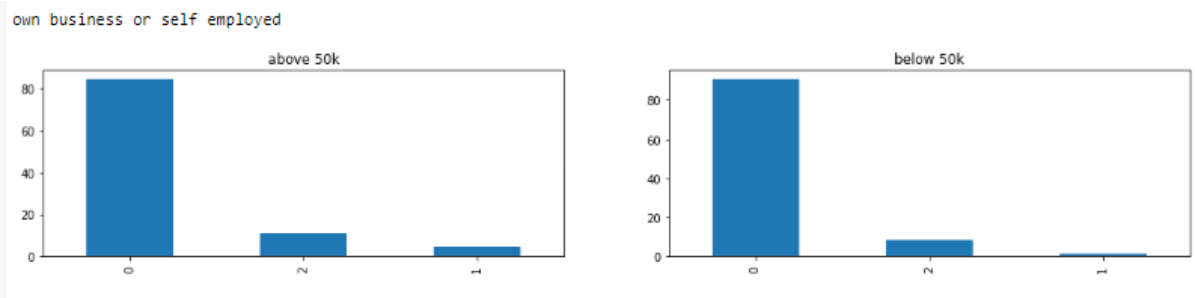


Figure 16 : Countplot of Business Owner w.r.t. Target

Weeks worked in year

Figure 17, From the below chart we can say that- In data weeks worked in year is a numerical feature but there is repetitions of all value many times as there are only 53 unique entries in this feature so we can convert weeks worked in year in categorical features, in the data 46.88 % of people worked for 0 weeks and almost 36% of people work for 52 weeks in year:

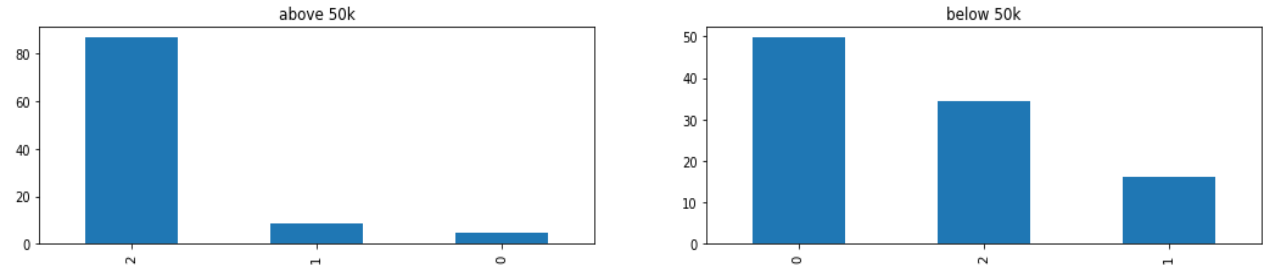


Figure 17 : Countplor of Weeks Worked in year w.r.t. Target

Citizenship_combined:

This is a new feature created by combining the 'country of birth father','country of birth mother', 'country of birth self' and 'citizenship'. The below figure 18 shows the barplot of the same. It is very much clear from the census year that most of the citizens are native because in 1994-95 globalisation was not that intense. Thus we get to see than 70% of data records are native citizens of USA.

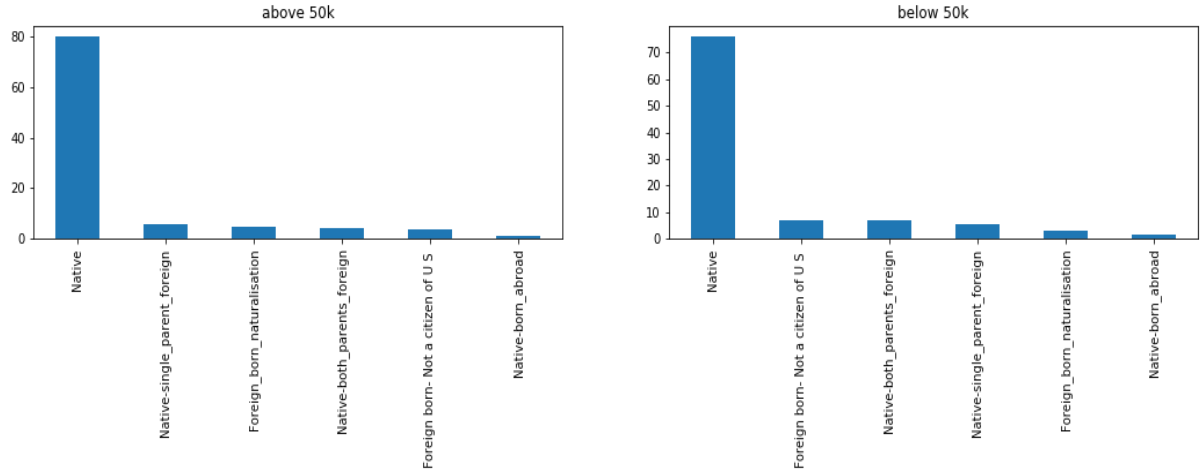


Figure 18 : Bar plot of Citizenship Combined w.r.t. Target

Capital Gains & Capital losses:

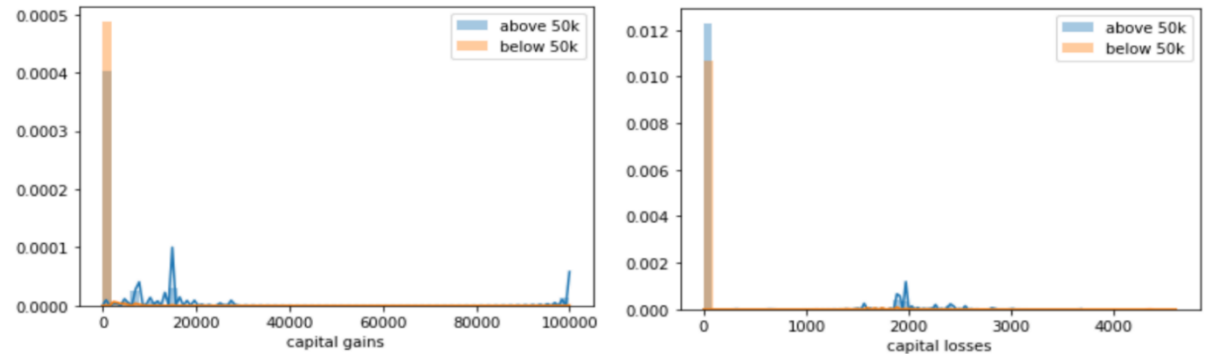


Figure 19 : Distribution plot of Capital Gains and Losses

Basic observation: People having income above 50k seem to have more capital gains or losses, it shows us a trend that this category tend to invest more.

Observation: After doing basic distribution on capital gains and losses, the data is highly discrete and there is a possibilty of outliers, hence we'll do feature engineering and combine capital gains with capital losses to make a new feature capital and check if that will give a better distribution

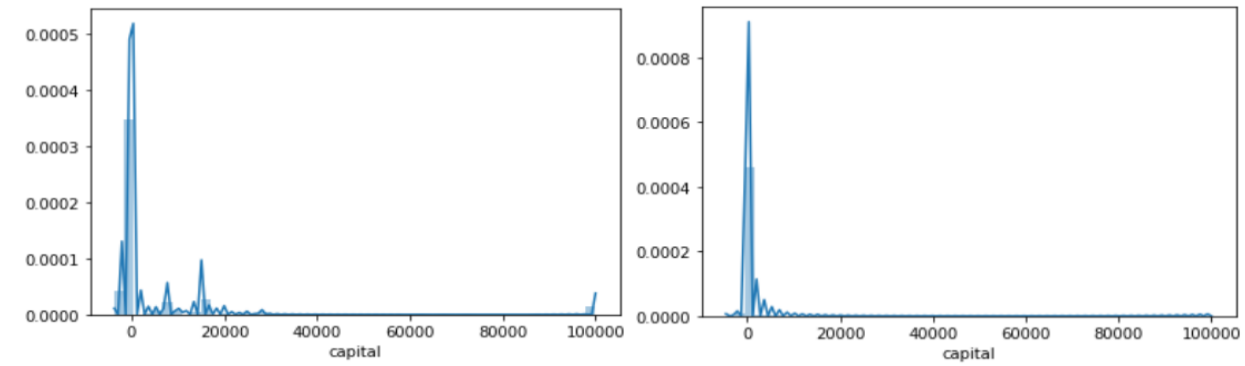


Figure 20 : Distribution of Capital (New Feature) w.r.t. Target

and better variance in with respect to the target variable.

We can apply transformation on numeric data or even try to remove outliers but for now we can see from visualisation that above 50k people have higher capital margins compared to less than 50k.

Dividends from stocks:

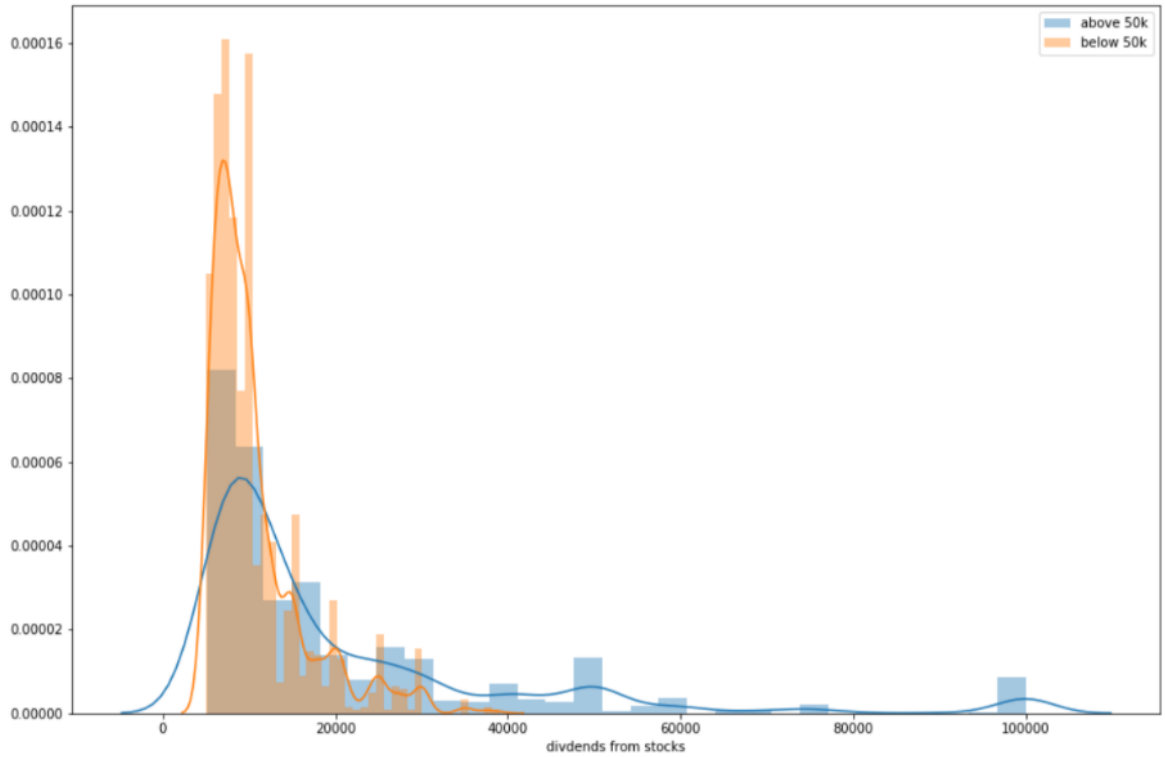


Figure 21 : Distribution plot of Dividends from Stocks

People with above 50k salary have more dividend gains and vice-versa. As it is a binary classification problem the outlier shouldn't be a problem but for safer side, the numeric data will be transformed because of high overall skewness.

Sex:

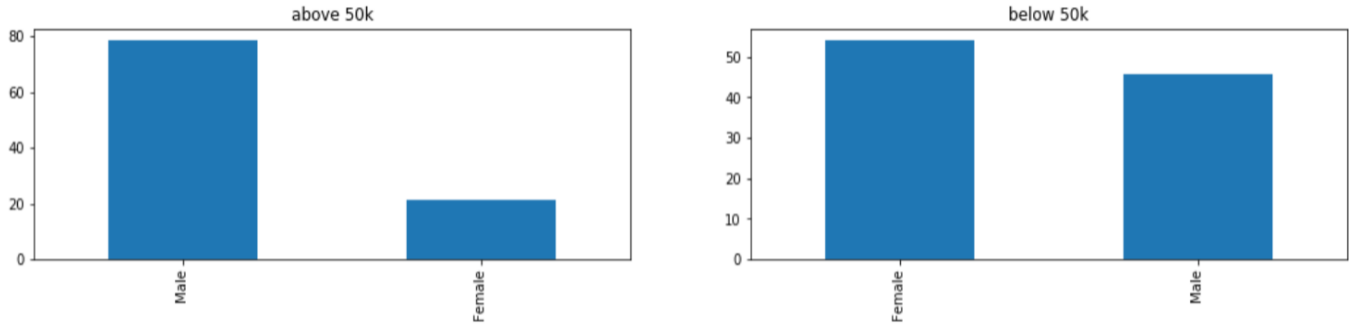


Figure 22 : Count plot of percentage of Sex w.r.t. Target

There are two categories male and female. From above Figure 22, we can see that there is more percentage of males in above 50k and more females in less than 50k. Definitely an important feature.

Detailed household summary in household:

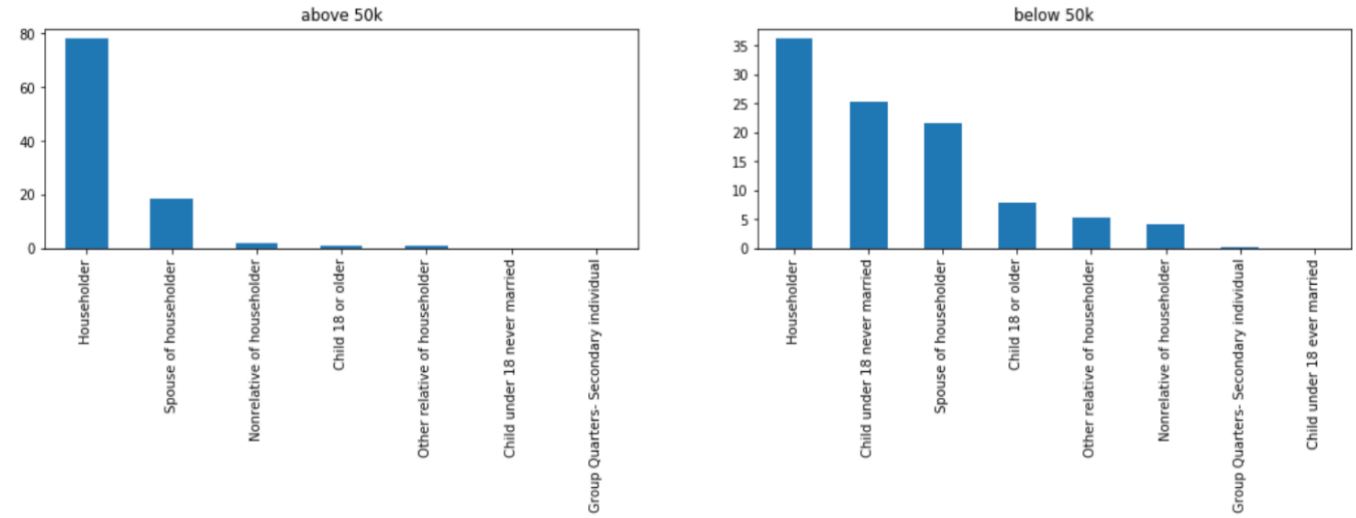


Figure 23 : Countplot of percentage of Detailed Household Summary w.r.t. Target

'detailed household and family stat' and 'detailed household summary in household' provide the same data. It depicts the role of instance in the family. One feature seems to be derived from another feature by combining similar categories hence *detailed household and family stat* will be dropped and *detailed household summary* will be kept and further category clubbing will be tried if similar categories found.

Marital Status:

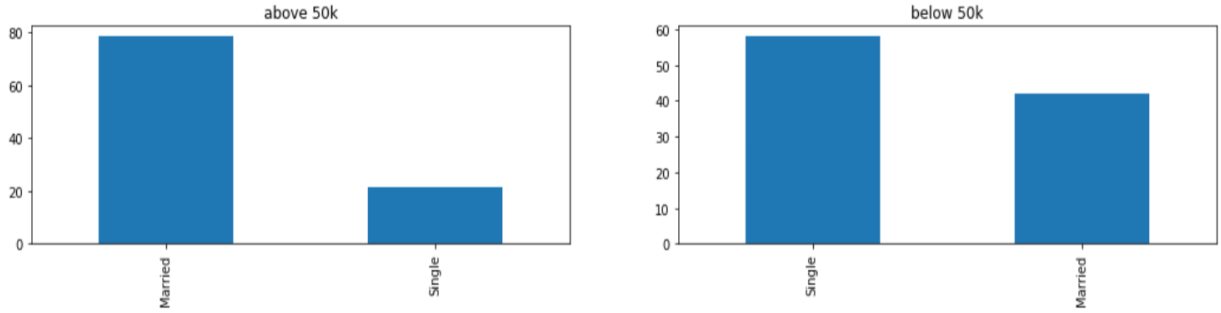


Figure 24 : Countplot of percentage of Marital Status w.r.t. Target

Married and single people as widowed, divorced, never married and separated represent single people only

Employment Status:

Adding a new feature 'Employment Status' instead of 'reason of unemployment'. we created this using feature engineering done on features like reason for unemployment, owns businees or self employed, major occupation code, age,. By doing this we got most of the data which are missing shown as 'Not in universe'.

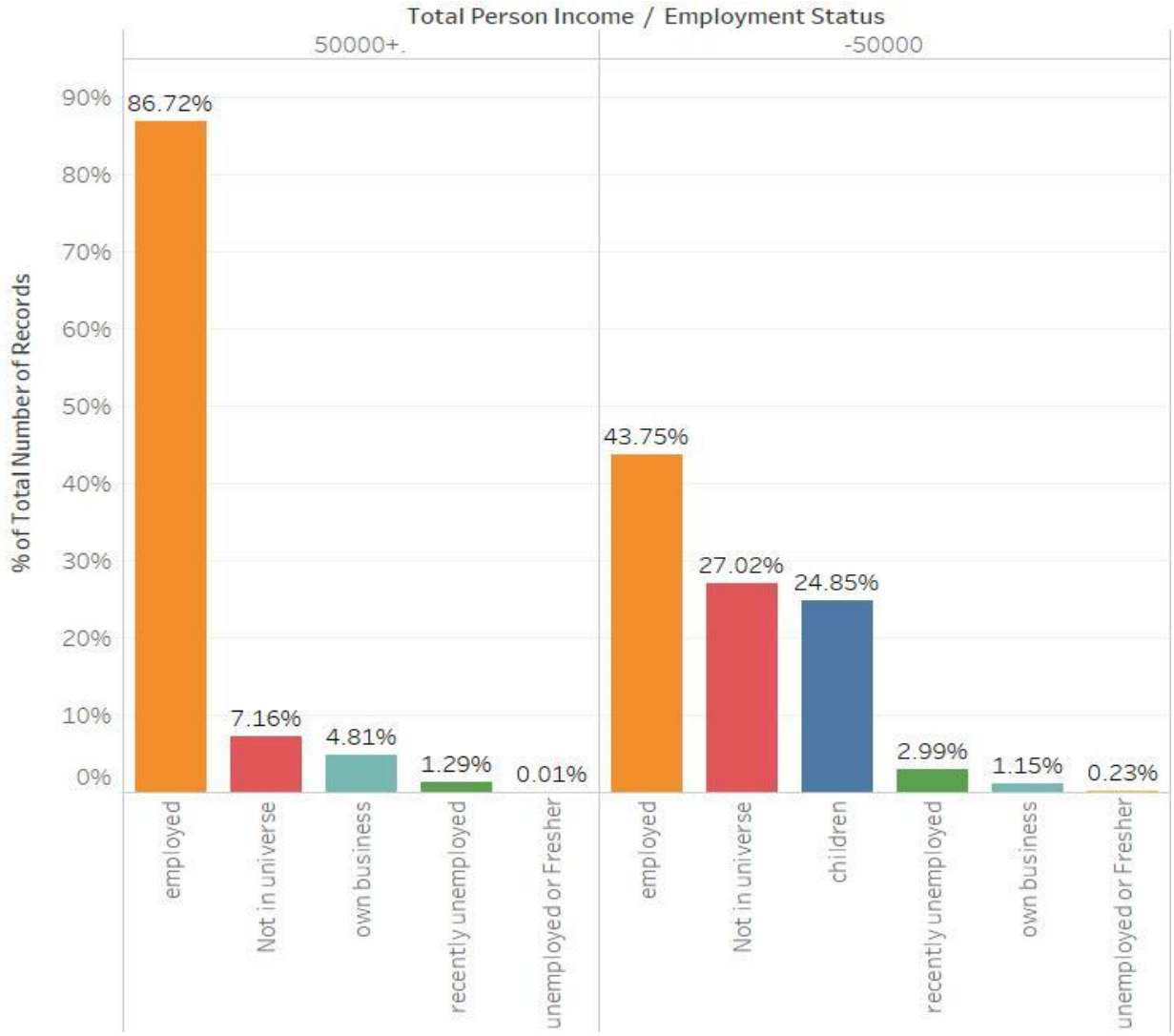


Figure 25 : Countplot of percentage of Employment Status w.r.t. Target

Observations in above figure 25: we can see a clear difference between two target bins now as children is present only in below 50k bin and 86.72% are employed in above 50 k bin whereas half of it only 43.75% are employed in below 50k bin. Also Children class is only present in Below 50k bin.

Race:

Race column contains White', ' Asian or Pacific Islander', ' American Indian Aleut or Eskimo', ' Black', ' Other' we can bucket all the people other than 'white' and 'black' people as minority.

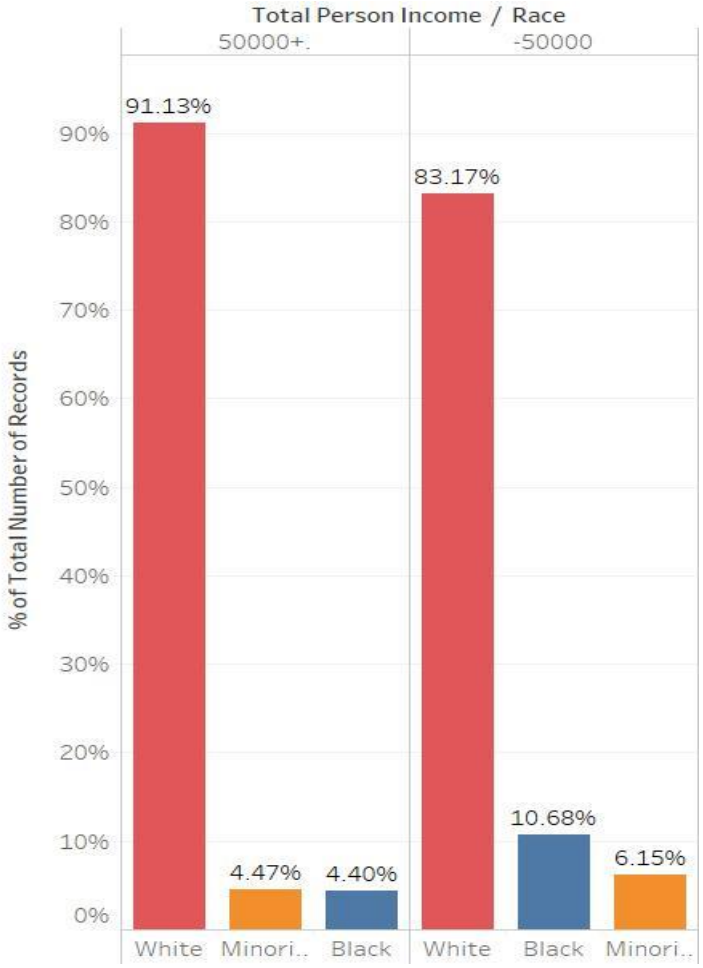


Figure 26 : Countplot of Percentage of Race (After Bucketing) w.r.t. Target

Observation: Even after Bucketing there's no significant improvement although black and minority are more in terms of percentage in lower income group.

Statistical Significance

After doing the feature engineering, we have been left with 3 Continuous and 16 Categorical independent variables. As our target variable is categorical, to find the p-value and make statistical inferences out of it we have to perform Anova Test on Continuous independent variables and Chi-square test on categorical independent variables.

Anova Test Results:

Following table represents p-value with respect each variable after Anova test:

Table 2 : Annova Test Results

Independent Variable	P-value	Test Result
wage per hour	0 < 0.05	Reject the null hypothesis
capital	2.23873365824334e-310 < 0.05	Reject the null hypothesis
dividends from stocks	0 < 0.05	Reject the null hypothesis

Chi-square Test Results:

Following table represents p-value with respect each variable after Chi-square test:

Table 3 : Chi-square Test Results

Independent Variable	P-value	Test Results
industry code	0 < 0.05	Reject the null hypothesis
occupation code	0 < 0.05	Reject the null hypothesis
class of worker	0 < 0.05	Reject the null hypothesis
marital status	0 < 0.05	Reject the null hypothesis
race	5.463433148269557e-126 < 0.05	Reject the null hypothesis
sex	0 < 0.05	Reject the null hypothesis
detailed household summary in household	0 < 0.05	Reject the null hypothesis
age	0 < 0.05	Reject the null hypothesis

weeks worked in year	0 < 0.05	Reject the null hypothesis
education	0 < 0.05	Reject the null hypothesis
full or part time employment stat	0 < 0.05	Reject the null hypothesis
tax filer status	0 < 0.05	Reject the null hypothesis
own business or self employed	4.17336860221902e-295 < 0.05	Reject the null hypothesis
Veteran's benefits	0 < 0.05	Reject the null hypothesis
employment status	0 < 0.05	Reject the null hypothesis
num persons worked for employer	0 < 0.05	Reject the null hypothesis
Citizenship_combined	1.6303798973841448e-90 < 0.05	Reject the null hypothesis

As the p-values for all the variables is below 0.05 cut-off, we reject the null hypothesis and can conclude that we have enough evidence to prove that all Independent variables are statistically different than dependent variable.

Feature Selection & Model Building

Feature selection is the process of selecting a subset of relevant attributes to be used in making the model in machine learning. Effective feature selection eliminates redundant variables and keeps only the best subset of predictors in the model which also gives shorter training times. Besides this, it avoids the curse of dimensionality and enhance generalization by reducing overfitting.

In this project, feature selection techniques are applied to improve the classification performance and/or scalability of the system. Thus, we aim to investigate if better or similar classification performance can be achieved with a smaller number of features. For feature ranking, instead of wrapper algorithms that require a learning algorithm to be used and consequently can result in reduced feature sets specific to that classifier, filter-based algorithms are tested in which no classification algorithm is used. Correlation and Probability Attribute Evaluation, Recursive feature Elimination and Minimum Redundancy Maximum Relevance Filters were used in our experiments. In mRMR algorithm, the aim is to maximize the relevance between the selected set of features and class variable while avoiding the redundancy among the selected features. Thus, maximum classification accuracy is aimed to be obtained with minimal subset of features.

Besides, considering the real-time usage of the proposed system, achieving better or similar classification performance with less number of features will improve the scalability of the system since less number of features will be kept track during the session.

Classification Results:

One of the purposes of this project is to get the predictions of even the minority class which is above 50k correctly. The dataset is fed to Logistic Regression, Random Forest, Light Gradient Boosting classifiers and other boosting methods using fivefold cross validations. The Accuracy, Recall, Precision, Bias Error and Variance Error and F1-Score are presented for each classifier. Most important metric is f1 score because it balances out recall and precision. The reason we are focusing both on recall and precision is we want to increase the True Positive rate with respect to incorrect predictions of both majority as minority and minority as majority. Both are affecting the wrong prediction of minority class.

Results on class imbalanced dataset:

Table 4 below show the results obtained with various algorithms. The results show that Light Gradient Boosting (LGBM) gives the highest f1 score on test set. However, a class imbalance

problem arises since the number of negative class instances in the data set is much higher than that of the positive class instances, and the imbalanced success rates on positive (TPR) and negative (TNR) samples show that the classifiers tend to label the test samples as the majority class. This class imbalance problem is a natural situation for the problem since most of the citizens earn low annual income.

Table 4 : Results of various Models applied on Imbalanced Dataset

Algorithm	Test_accuracy	Precision	Recall	Bias_Error	Variance_Error	F1	ROC
LGBM	95.58	0.75	0.43	0.0442	0.11	0.76	0.95
RF	95.09	0.65	0.44	0.0491	0.12	0.75	0.92
GB	95.41	0.74	0.41	0.0459	0.11	0.75	0.91
KNN	94.63	0.6	0.4	0.0537	0.08	0.73	0.94
ADB	95.17	0.72	0.36	0.0483	0.1	0.73	0.94
XGB	95.33	0.76	0.36	0.0467	0.09	0.73	0.94
LR	94.13	0.62	0.14	0.0587	0.14	0.64	0.84

Results obtained with oversampling:

The results presented in the table 5 below, shows that the classifiers tend reduce the overall predictions, but the Recall value is better when compared to unsampled data. To deal with class imbalance problem, we use oversampling method, in which a uniform distribution over the classes is aimed to be achieved by adding more of the minority (positive class in our dataset) class instances. Since this dataset is created by selecting multiple instances of the minority class more than once, first oversampling the dataset and then dividing it into training and test sets may lead to biased results due to the possibility that the same minority class instance may be used both for training and test. For this reason, in our study, 30 percentage of the data set consisting of 85K samples is first left out for testing and the oversampling method is applied to the remaining 70 percentage of the samples.

Table 5 : Results of various Models applied on Oversampled Dataset

Algorithm	Test_accuracy	Precision	Recall	Bias_Error	Variance_Error	F1	ROC
RF	91.34	0.38	0.63	0.0866	1.45	0.71	0.91
ADB	91.96	0.4	0.6	0.0804	0.61	0.71	0.9
LGBM	88.7	0.33	0.79	0.113	0.96	0.7	0.93
XGB	88.34	0.32	0.77	0.1166	0.48	0.69	0.92
GB	85.52	0.28	0.85	0.1448	0.56	0.67	0.93
KNN	89.18	0.31	0.59	0.1082	0.51	0.67	0.89
LR	77.12	0.19	0.83	0.2288	0.37	0.59	0.88

The results obtained on the balanced dataset are shown in the table below. Since the number of samples belonging to positive and negative classes is equalized with oversampling, both accuracy and F1-score metrics can be used to evaluate the results.

Feature Selection and Model Interpretability:

The LGBM algorithm, which achieved the highest accuracy and F1-score, has been chosen to identify the most significant features affecting the salary of an American citizen. In this section, we apply feature selection to further improve the classification performance of LGBM classifier. Besides, considering the real-time usage of the proposed system, achieving better or similar classification performance with a smaller number of features will improve the scalability of the system since a smaller number of features will be kept track during future analysis.

Applied LGBM model on non-over-sampled data:

Table 6 : Feature Selection & Model Interpretability

Feature Selection	accuracy score	Roc-score	precision	recall	f1-score	bias-error	variance
RFE-logreg	0.9431	0.9168	0.61	0.24	0.66	0.0569	0.1253
RFE-RF	**0.9558**	**0.9497**	**0.75**	**0.43**	**0.76**	0.0442	0.1304
mRMR-15	0.9466	0.9304	0.65	0.3	0.69	0.0534	0.1144
mRMR-10	0.9431	0.9169	0.61	0.24	0.66	0.0569	0.0832

From above table 6, recursive feature Elimination using Random Forest provides a better model using selected features on LGBM Classifier.

Insights inferred through odds-Ratio using Logistic Regression:

Odds ratio = exp(coef(LR model))
Probability = Odds/(1+Odds)

The Odds ratio and probability of some of the variables, based on their practical significance have been listed below in table 7:

Table 7 : Odds ratio and probability of the variables

VARIABLE	CO-EF	ODDS RATIO	PROBABILITY	CORRELATION
industry code	0.27252	1.027627	50.68%	**0.13**
occupation code	-0.198001	0.820369	45.06%	0.09
class of worker	0.068370	1.070761	51.70%	**-0.10**
marital status	-0.278214	0.757135	43.08%	**-0.18**

race	0.206407	1.229254	55.14%	0.03
sex	1.390707	4.017688	80.07%	**0.16**
detailed household summary in household	0.049506	1.050752	51.23%	0.07
full or part time employment stat	-0.020326	0.979879	49.49%	0.02
tax filer status	-0.076788	0.926086	48.08%	-**0.14**
employment status	0.411870	1.509638	60.15%	**0.16**
Citizenship combined	0.057998	1.059713	51.44%	0.00
wage per hour	-0.000172	0.999828	49.99%	0.02
capital	0.000108	1.000108	50.00%	**0.23**
dividends from stocks	0.000218	1.000218	50.00%	**0.18**
age	0.497757	1.645028	62.19%	**0.13**
weeks worked in year	1.343116	3.830963	79.30%	**0.26**
education	0.512799	1.669959	62.54%	**0.25**
own business or self employed	-0.047277	0.953823	48.81%	0.04
Veteran's benefits	-0.061611	0.940249	48.46%	**0.14**
number persons worked for employer	0.171989	1.187665	54.28%	**0.22**

Inferences:

- If the sex of person is Male (1), then 80% of probability that person is earning more than 50K.
- If the person works for higher number of weeks per year, than the probability of the person earning more than 50K increases by 79.3% at each level.
- Employment status also has positive coefficient which suggest above 50k with better employment status (ordinal data).

Variable Importance Plot for Tree Based Algorithms

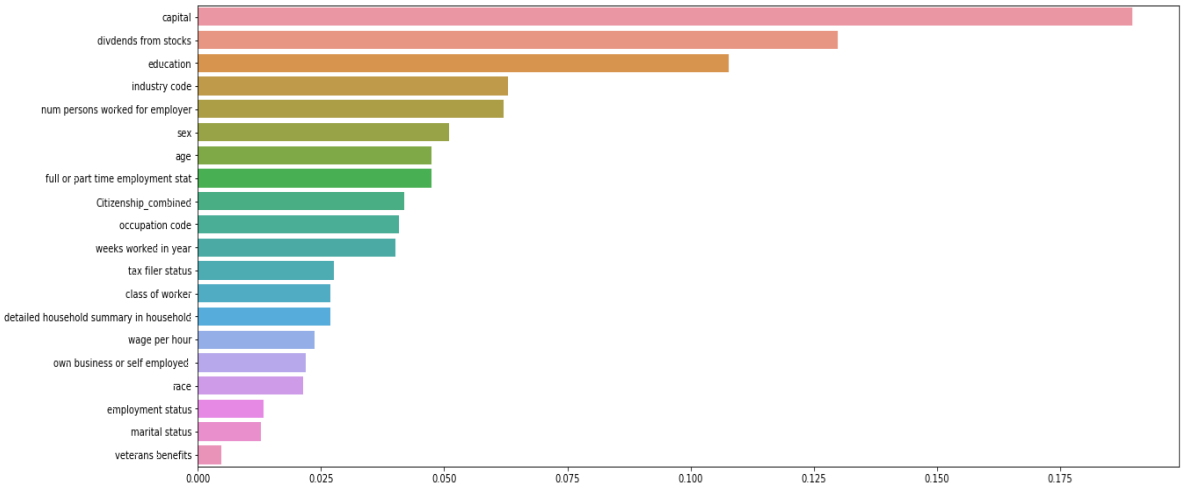


Figure 27 : Plot of Feature Importance from Decision Tree Model

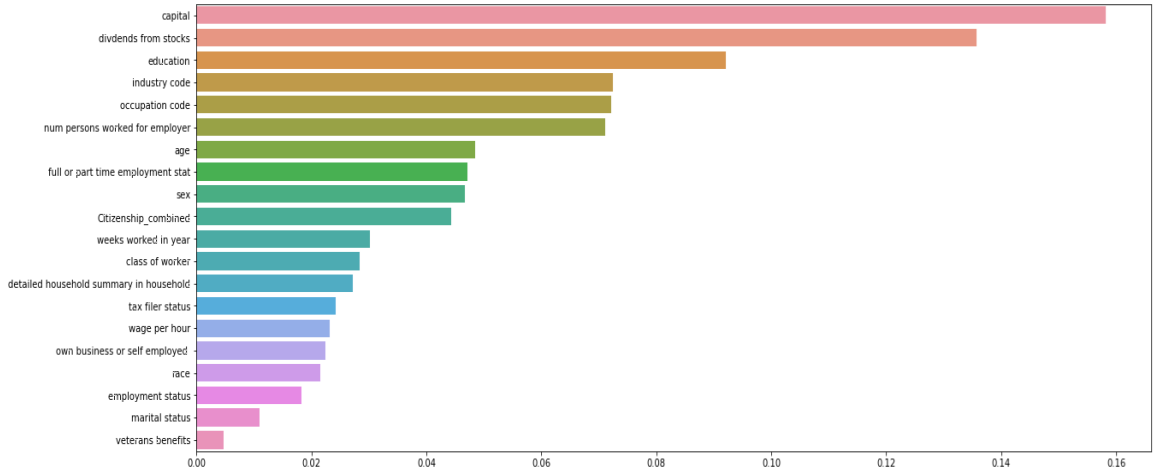


Figure 28 : Plot of Feature Importance from Random Forest Model

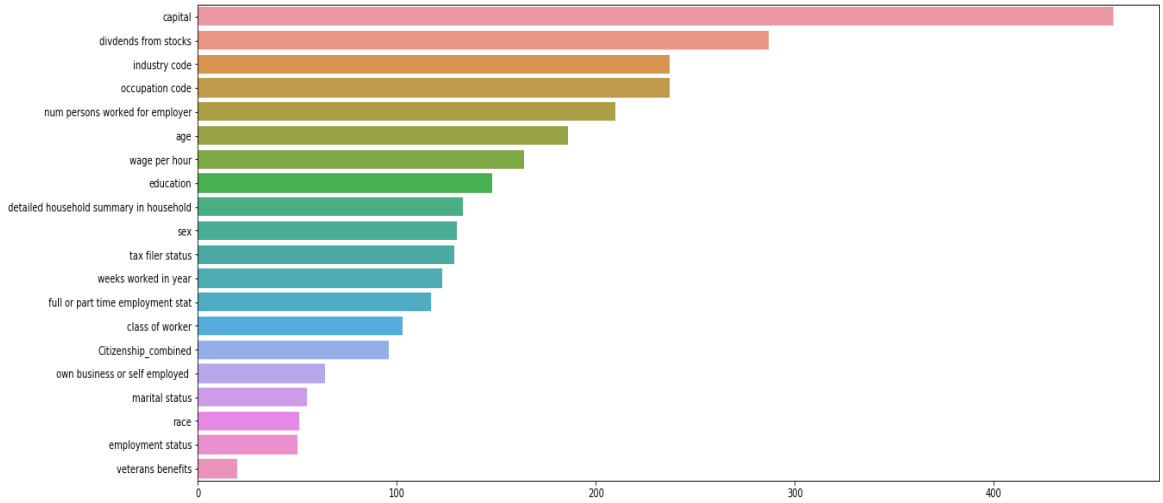


Figure 29 : Plot of Feature Importance from Gradient Boosting Model

Figures 27, 28 and 29 shows the important features from the models Decision tree, Random Forest and Gradient Boosting respectively. From these plots we can say that following are the important 12 features which are most important in all the before mentioned plots:

- Capital
- Dividends from Stocks
- Education
- Industry Code
- Occupation Code
- Num of persons worked for employer
- Age
- Full or Part time employment Stats
- Citizenship Combined
- Sex
- Weeks Worked in year
- Tax Filer Stats

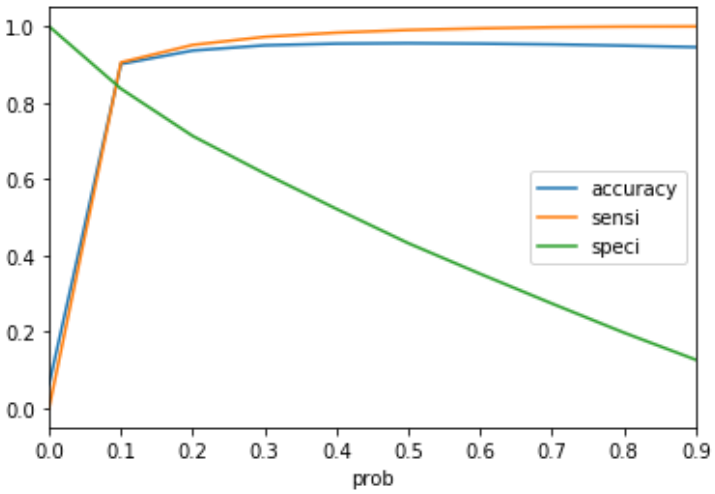
Thresholding

Thresholding is the process of adjusting the cutoff probability of an instance being classified as 0 or 1

Comparing results on thresholding the probability of LGBM model with unsampled data

Threshold	Precision	Recall	F1
0.5	0.75	0.43	0.76
0.4	0.68	0.52	0.78
0.3	0.60	0.61	0.79
0.2	0.49	0.71	0.77

Significant tradeoff results at 0.3 cutoff. Hence the final model considered by us is LGBM Classifier working upon un-sampled data and threshold at 0.3 probability value.



	precision	recall	f1-score	support
0	0.97	0.97	0.97	93576
1	0.60	0.61	0.61	6186
accuracy			0.95	99762
macro avg	0.79	0.79	0.79	99762
weighted avg	0.95	0.95	0.95	99762

Conclusions

In this project, we aim to analyze the different attributes affecting the salary of an American citizen and build a predictive model for the same. We use US census data to perform the experiments. In order to predict the salary bracket of the person we used data such as education, marital status, gender, industry, benefits, capital returns etc. as input to machine learning algorithms. Oversampling and feature selection pre-processing techniques are applied to improve the success rates and scalability of the algorithms. The best results are achieved with a Light Gradient Boosting algorithm calculated using resilient backpropagation with weight backtracking. Our findings support the argument that the features drawn from census are important for the prediction of target variable.

Although we have done exhaustive EDA and Feature Engineering to reduce redundant information, we use feature selection for better understanding of mathematical relationship between features and target variable. Therefore, we apply a feature ranking method. Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output. The findings show that removing some less important features like citizenship and owning of business results in a more accurate and scalable system. Considering the real time usage of the proposed system, achieving better or similar classification performance with minimal subset of features is an important factor for better ML modelling since a smaller number of features will be kept track during the data collection.

From Tables 4 and 5 it is clear that oversampling is giving poor precision scores resulting into poor Bias and variance. Oversampling is supposed to give us the better model but it is worsening the situation. This is the trade-off we have accept here.

As per our Trade-off, following are the results for the un-sampled data with LGBM model:

- Precision = 0.6 , Recall = 0.61 , F1-score = 0.79
- Accuracy = 95%

Recommendations and Actionable Insights

The data has a lot of story to tell. Some Actionable insights which various government institutions can use to improve the earning disparity and many other goals which can be achieved from the inferences.

1. As we can see there are more percentage of males in above 50k and more females in less than 50k category. Govt can use this insight to empower women and make reforms to balance out the income disparity.
2. Racial discrimination is less than expected in above 50k category which was prevalent in US in the past.
3. People owning business tend to have better income but it's uncertain because the income is divided in just two categories. Government can promote people to start with their own business by adjusting their laws, which can help in increase of per capita income.
4. Government can encourage people to invest because it is definitely a good sign with respect to better income. Be it stock market investment or capital investment(property), it affects the annual income of a person.
5. Government should promote higher level of education as it directly affects annual income of a person.
6. Industry code can be used to understand which industry is booming and giving above 50k income to the employees. Such data can help young students to select their field of study.
7. Personal financial planning can be done according to the data. The age distribution tells that 25-40 is the age where chances of earning more that 50k is more.

References & Bibliography

- Current Population Survey, March 1994 & March 1995 Technical Documentation. All the census data is available from the Administrative and Customer Services Division, Customer Services, Bureau of the Census, Washington, DC 20233. Source locations:

    1. 1994 Survey: https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar94.pdf
    2. 1995 Survey: https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar95.pdf

- To get to know about US government policies regarding legal working age, Demographical areas of US, Marriage & Divorce related queries we referred official website of US Census Bureau https://www.census.gov/.

- To get to know about labor and industry laws, employment and unemployment related categories and standard working hours for part time and full time referred census's official website. https://www.census.gov/.

- Also referred Wikipedia for Veteran related queries, also getting to know about the educational system of US and migration related queries. https://www.wikipedia.org/

Appendix

Column information

Age (Continuous)	91 distinct values for attribute ranging from 0 to 91
class of worker (nominal)	9 distinct values for attribute. This refers to the broad classification of the person's employer.
detailed industry recode (nominal)	52 distinct values for attribute ranging from 0 to 51. There are 236 categories for the employed, with 1 additional category for the experienced unemployed. These categories are aggregated into 52 detailed groups / recodes.
detailed occupation recode (nominal)	47 distinct values for attribute ranging from 0 to 46. There are 500 categories for the employed with 1 additional category for the experienced unemployed. These categories are aggregated into 46 detailed groups / recodes.
education (nominal)	17 distinct values for attribute. This represents educations had by every individuals.
wage per hour (continuous)	1240 distinct values for attribute. Ranges from 0 to 9999. This basically represents hourly income of every individual.
enroll in edu inst last wk(nominal)	3 distinct values for attribute. Represents if an individual has enrolled in the educational institute in last week of the Survey.
marital stat (nominal)	7 distinct values for attribute. Marital Status of an individual till survey.
major industry code (nominal)	24 distinct values for attribute. There are 236 categories for the employed, with 1 additional category for the experienced unemployed. These categories are aggregated into 24 detailed groups / recodes.
major occupation code (nominal)	15 distinct values for attribute. There are 500 categories for the employed with 1 additional category for the experienced unemployed. These categories are aggregated into 15 detailed groups / recodes.
race (nominal)	5 distinct values for attribute.
Hispanic origin (nominal)	10 distinct values for attribute.
sex (nominal)	2 distinct values for attribute
Member of a labor union (nominal)	6 distinct values for attribute. a member of a labor union or of an employee association similar to a union
Reason for unemployment (nominal)	6 distinct values for attribute. Reason for unemployment. Unemployed persons are those civilians who, during the survey week, have no employment but are available for work.
Full or part time employment stat (nominal)	8 distinct values for attribute
Capital gains (continuous)	132 distinct values for attribute. Amount of capital gains in dollars.
capital losses (continuous)	132 distinct values for attribute. Amount of capital losses in dollars.
dividends from stocks (continuous)	1478 distinct values for attribute. How much did an individual receive in dividends from stocks (mutual funds)?
tax filer stat (nominal)	6 distinct values for attribute. Tax filing status of an individual.

region of previous residence (nominal)	6 distinct values for attribute. Region of previous residence from South, west, Northeast, mid-west or abroad.
state of previous residence (nominal)	51 distinct values for attribute
detailed household and family stat (nominal)	38 distinct values for attribute. Details about family members, child, parents, grand-parents.
detailed household summary in household (nominal)	8 distinct values for attribute. Summary of feature 'detailed household and family stat'
instance weight (continuous)	The instance weight indicates the number of people in the population that each record represents due to stratified sampling. To do real analysis and derive conclusions, this field must be used. This attribute should *not* be used in the classifiers, so it is set to "ignore" in this file.
migration code-change in msa (nominal)	10 distinct values for attribute. MSA- metropolitian statistical code.
migration code-change in reg (nominal)	9 distinct values for attribute. Migration from one region to another
migration code-move within reg (nominal)	10 distinct values for attribute. Migration within a region.
live in this house 1 year ago (nominal)	3 distinct values for attribute
migration prev res in sunbelt (nominal)	4 distinct values for attribute
num persons worked for employer (continuous)	7 distinct values for attribute ranging from 0 to 6. Counting all locations where this employer operates, what is the total number of persons who work for record's employer?
family members under 18 (nominal)	5 distinct values for attribute. This excludes reference person and spouse if under 18.
country of birth father (nominal)	43 distinct values for attribute. Name of a country
country of birth mother (nominal)	43 distinct values for attribute. Name of a country
country of birth self (nominal)	43 distinct values for attribute. Name of a country
citizenship (nominal)	5 distinct values for attribute. Allocated citizenship to an individual.
own business or self employed (nominal)	3 distinct values for attribute. Represents data of self-employed or business owner from the record.
fill inc questionnaire for veteran's admin (nominal)	3 distinct values for attribute. This is only for Armed Force retired individual.
veterans benefits (nominal)	3 distinct values for attribute. Record of individual receiving benefits after participating in war.

weeks worked in year (continuous)	53 distinct values for attribute ranging from 0 to 52.
year (nominal)	2 distinct values for attribute. Whether record belongs to 1994 or 1995 survey.
Total Person Income (nominal)	2 distinct values for attribute. Whether record bin of 'below $50k yearly income' or 'above $50k yearly income'