Predict the income of adult from dataset of census income

Yujie Zhang, Master of Professional Study

University of Auckland

Abstract

With the improvement of monetary world broadly, SDG distributed the objective of advancing supportable financial development and gainful work for all in the world. The most effective method to accomplish more elevated amounts of profitability in different regions by giving all the more nice activity is one of the objectives for this objective. The article targets foreseeing the salary of a grown-up as per the aftereffect of machine learning on the dataset of registration pay. The strategy for the looking into incorporates a few stages like business understanding, information understanding, information planning, information mining technique choice, information mining calculation determination, information mining, elucidation. Python is chosen as the product suite for this undertaking. Ends and further work are given at last after the examination as indicated by the aftereffect of break down.

Keywords: Machine Learning, python

Introduction

Motivation

Sustainable Development Goals can be implemented by means which are generated by the growth of inclusive and sustainable economic growth which is also drive progress of it.

Nowadays the practical problem of relatively low productivity and high level of unemployment still can not be resolved. Meanwhile, the global economy is growing at a slower rate. To increase the opportunities of employment, especially for young people, more progress is needed to increase opportunities of employment, reduce employment which is informal with the pay gap of gender and also secure the environments of working for decent work, more progress is needed.

The motivations of this project are to analyse the dataset to find the most important factors which are able to effect and predict the income of adult. The research problems of this project are to answer the following questions:

What is the most important feature which affect the income of adult?

What is the relationship between income of adult with the other features?

How to predict if the income of adult would be larger than 50K or less than it?

Which is the best model to do the prediction and the accuracy of it?

Research objectives

To define a technical solution to this business problem, all things need to be kept concrete. So the research objectives of the project need be translated into data mining term. The goals for the initial study to be completed are:

- Use historical information about previous income of adults to generate multiple models of classification to perform comparisions between different models and find the pattern of the data. Also, the importance and relationship between factors would be provided by this project.
- Select correct method and algorithm to predict if the income for an adult would be larger or less than 50K per year.
- The evaluation of the model will be performed in different way including logarithms loss, confusion matrix and classification report. The model with high performance will be selected as the best model in the end. The basic line to select the model is that the accuaracy is larger than 70% and the f1 score of the model is larger than 0.7.

Literature

A Comparative Study of Performances of Various Classification Algorithms for Predicting Salary Classes of Employees.

S. Chakraborti (2014) used the public database (UCI census dataset) which have most of the attributes available for a segment of population for salary prediction. This paper investigates the comparative performance of a portion of these standard classification algorithms on a notable dataset, namely census dataset available from UCI. The fundamental reason is to develop clear comprehension about the applicability of these algorithms for salary prediction of employees of an organization and their comparative performance figures. This investigation found that among five classification algorithms, decision tree and Bayesian belief network performs superior to other three algorithms like naïve Bayes, support vector machine and neural network. The software utilised for running these algorithms is WEKA which is a notable university tool for machine learning.

Random Forest for Salary Prediction System to Improve Students' Motivation

Khongchai, P. And Songmuang, P. (2016) generated a salary prediction model for graduate students using a data mining technique to generate for individuals with similar training attributes. An experiment was conducted to compare the two techniques of data mining including Decision Trees ID3, C4.5 and Random Forest so that they can determine the most suitable method to predict the salary of adults , tuned with important parameters to improve the accuracy of the results. Random Forest gave the best precision at 90.50%, while Decision Trees ID3 and C4.5 returned lower accuracies at 61.37% and 73.96%, separately for 13,541 records of graduate students using a 10-fold cross-validation method. Random Forest generated the best efficiency model for salary prediction. Results indicated that the system was effective in boosting students' motivation for studying, and also gave them a positive future viewpoint. It also suggested that the students were satisfied with the implemented system because it was easy to use, and the prediction results were simple to understand even without previous background knowledge of statistical.

Salary Predictor System for Thailand Labour Workforce using Deep Learning

Viroonluecha, P. And Kaewkiriya, T (2018) built a Salary Predictor System to predict monthly salary of employees in Thailand using the Deep Learning approach, which has rapidly increased distinguish attention in machine learning field. The dataset has been assembled from the famous job search website which has

more than 1.7 million clients. Individual information from the first five months of 2018 is applied to the analysis and develop this model. They contrasted the performance with related algorithms such as Random Forest and Gradient Boost Trees. The feature selection methods were applied to Deep Learning in the wake of comparing. As a result of combining the feature selection with Deep Learning, the optimal outcome was 0.462 in R-squared and the rapid runtime is 15.37 seconds.

Methodology

Machine Learning. Supervised, unsupervised and reinforcement are three main methods of machine learning. Supervised and unsupervised are mainly used to deal with projects of machine learning. Reinforcement learning is more powerful and complex to apply for problems.Figure 1 shows the structure of methods of matching learning.(Madhu Sanjeevi, 2017)

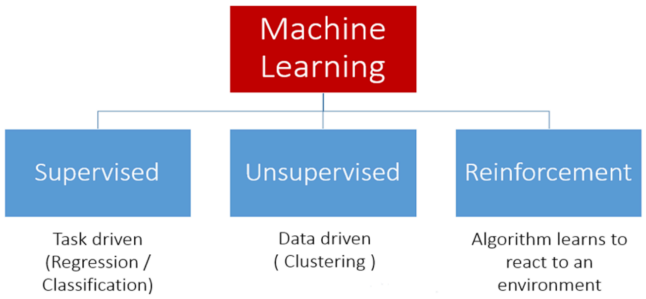

Figure 1. Methods of Matching Learning

Supervised learning is the machine learning task of learning a capacity that maps a contribution to a yield dependent on model information yield pairs. It gathers a capacity from labeled training data consisting of a set of training examples. In directed adapting, every model is a pair consisting of an information object (regularly a vector) and an ideal yield esteem (likewise called the supervisory signal). A regulated learning calculation breaks down the preparation information and produces an induced capacity, which can be used for mapping new models. An ideal situation will consider the calculation to accurately decide the class marks for concealed cases. This requires the taking in calculation to sum up from the preparation information to concealed circumstances in a "sensible" manner. Figure 2 shows the main types of supervised learning. (Madhu Sanjeevi, 2017)

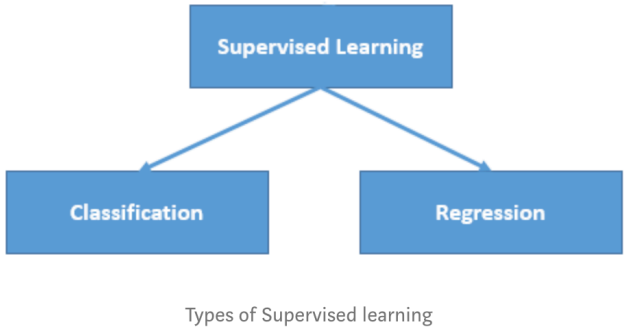

Figure 2. Types of Supervised learning

Regression: This is a type of problem where we need to predict the *continuous-response* value

Classification: This is a type of problem where we predict the *categorical response* value where the data can be separated into specific "**classes**". Classification and regression trees are machine-learning techniques which are used to develop forecast models from information. The calculation recursively parcel the information space and fit a basic expectation model inside each segment. A chart of choice tree will be produced by the calculation. Classification trees and regression trees are two fundamental segments of the calculation. A limited number of unordered qualities, with forecast mistake estimated as far as misclassification cost are taken as contributions of arrangement trees. Consistent or requested discrete qualities, with forecast mistake commonly estimated by the squared distinction between the watched and anticipated qualities are taken as contributions of relapse trees.(Loh, 2011).

Classification, which is one of the most important learning models in data mining, aims at building a model to predict future behaviours through classifying datasets into multiple predefined classes based on certain criteria. Neural networks, decision trees and if-then-else rules are common tools which are used for classification (Ngai, 2009).

Unsupervised learning

Unsupervised learning may be a sort of machine learning algorithm utilized to draw deductions from datasets comprising of information data without checked responses. The foremost broadly recognized unaided learning strategy is cluster examination, which is utilized for exploratory data examination to discover concealed illustrations or gathering in data. The bunches are illustrated employing an extent of resemblance which is characterized upon estimations, for case, Euclidean or probabilistic partition. The preparing information does not include Targets here so we don't tell the framework where to go , the framework should get it itself from the information we allow. Figure 20 appears the method of unsupervised learning. Figure 3 shows the process of unsupervised learning. (Madhu Sanjeevi, 2017)

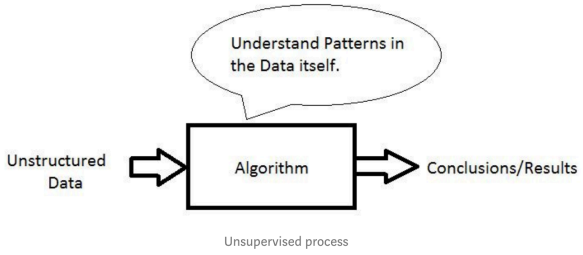

Figure 3.Unsupervised process

Logistic Regression

Logistic regression is the proper relapse examination to direct when the reliant variable is dichotomous (binary). It is used when the dependent target is categorical. Like all regression investigations, logistic regression is a prescient analysis. It is utilized to depict information and to clarify the connection between one ward twofold factor and at least one ostensible, ordinal, interim or proportion level autonomous factors. In logistic regression, a solitary result variable Yi (I = 1,..., n) pursues a Bernoulli likelihood function that takes on the worth 1 with likelihood πi and 0 with likelihood 1 − πi . At that point πi changes over the perceptions as a backwards strategic capacity of a vector xi , which incorporates a steady and k − 1 illustrative variables.(Statistical Solution, 2019)

The objective of logistic regression is to locate the best fitting (yet organically sensible) model to depict the connection between the dichotomous normal for intrigue (subordinate variable = reaction or result variable) and a lot of autonomous (indicator or informative) factors. Strategic relapse produces the coefficients (and its standard blunders and importance levels) of an equation to anticipate a logit transformation of the likelihood of essence of the normal for intrigue. (Medcalc, 2019)

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values. Figure 20 showed the following code to create a model of logistic regression. (Medcalc, 2019)

SVM

The SVM classifier is a run of the mill discriminative classifier. Unique in relation to generative classifier, it for the most part centers around how well they can isolate the positives from the negatives, and doesn't attempt to comprehend the fundamental data of the individual classes. The SVM classifier maps first the case x in a preparation set into a high dimensional space through a capacity Φ, at that point processes a choice capacity of the structure f(x) = + b by amplifying the separation between the arrangement of focuses Φ(x) to the hyperplane or set of hyperplanes parameterized by (w, b) while being steady on the preparation set. SVM is a hyperplane which can be utilized to isolate the positive information from the negative information with most extreme edge in the element space. The edge represents the separation between the hyperplane with the closest of the positive and negative information. (Hwanjo Yu, 2002). SVMs expand the edge (Winston phrasing: the 'road') around the isolating hyperplane. The choice capacity is completely indicated by a (normally exceptionally little) subset of preparing tests, the help vectors.This turns into a Quadratic programming issue that is anything but difficult to comprehend by standard strategies (M. E. Mavroforakis, 2006). Figure 21 shows the ideas of SVM.

A SVM finds the best isolating (maximal edge) hyperplane between the two classes of preparing tests in the element space, as it is appeared in Fig. 1. A direct discriminant capacity has the type of the straight practical , which relates to a hyperplane, separating the component space. On the off chance that, for a given example mapped in the component space to , the estimation of is a positive number, at that point the example has a place with the class named by the numeric worth ; else, it has a place with the class with worth.(Hwanjo Yu, 2002).

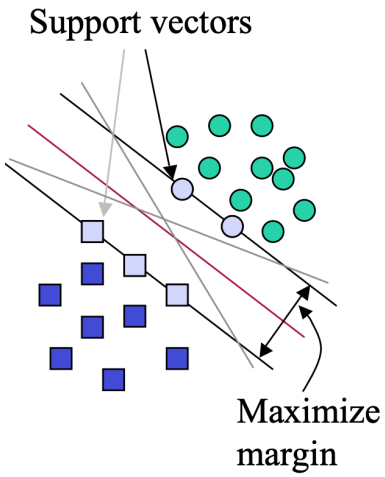

Figure 4. SVM Idea

Decision Tree

Decision trees are among the notable AI methods. A decision tree is made out of three essential components. A decision node which is indicating a test characteristic. An edge or a branch which is relating to the one of the conceivable trait esteems which means one of the test quality results. A leaf which is additionally named an answer hub, contains the class to which the item has a place (N. Amor, 2004). Figure 5 shows the idea of decision tree.

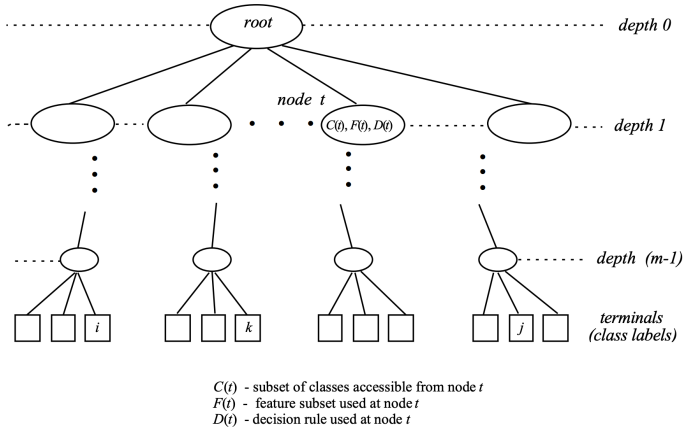

Figure 5. Decision Tree

In decision trees, two noteworthy stages ought to be guaranteed:

1. Building the tree. In view of a given preparing set, a choice tree is fabricated. It comprises of choosing for every choice hub the 'proper' test credit and furthermore to characterise the class marking each leaf.

2. Grouping. So as to group another example, we start by the foundation of the choice tree, at that point we test the characteristic determined by this hub. The aftereffect of this test permits to descend the tree limb with respect to the characteristic estimation of the given occurrence. This procedure will be rehashed until a leaf is experienced. The occasion is then being ordered in a similar class as the one describing the arrived at leaf (Shi, H, 2011). Figure 24 shows the code which is used to create a model of decision tree.

Random Forests

There are three main steps for the algorithm of random forest. 1. Draw ntree bootstrap tests from the first information. 2. For every one of the bootstrap tests, grow an unpruned arrangement or relapse tree, with the accompanying change: at every hub, as opposed to picking the best split among all indicators, arbitrarily test of the indicators and pick the best split from among those factors. 3. Anticipate new information by conglomerating the expectations of the ntree trees. A gauge of the blunder rate can be gotten, in light of the preparation information, by the accompanying: 1. At each bootstrap emphasis, anticipate the information not in the bootstrap test utilising the tree developed with the bootstrap test. 2. Total the OOB expectations. Calculate the blunder rate, and consider it the OOB gauge of mistake rate. The randomForest bundle alternatively delivers two extra snippets of data: a proportion of the significance of the indicator factors, and a proportion of the inside structure of the information (the vicinity of various information focuses to each other). Variable significance This is a troublesome idea to characterise by and large, on the grounds that the significance of a variable might be because of its association with different factors. The arbitrary timberland calculation assesses the significance of a variable by taking a gander at how much expectation mistake increments when (OOB) information for that variable is permuted while all others are left unaltered. The vital computations are done tree by tree as the irregular timberland is built. The instinct is that "comparative" perceptions ought to be in a similar terminal hubs more frequently than unique ones (Liaw A, 2002).

Naive Bayes

Bayes frameworks are one of the most by and large used graphical models to address and deal with questionable information. Bayes frameworks are dictated by two portions: - A graphical part made out of an organised non-cyclic diagram (DAG) where vertices address events and edges are relations between events. - A numerical part including in an assessment of different associations in the DAG by a prohibitive probability allocation of each centre with respect to its people. Honest Bayes are astoundingly clear Bayes frameworks

which are made out of DAGs with only one root centre (called parent), addressing the secretly centre, and a couple of youths, contrasting with watched centres, with the strong assumption of self-governance among child centre points concerning their parent. The portrayal is ensured by accepting the parent centre to be a disguised variable communicating to which class each article in the testing set should have a spot and adolescent centre points address different qualities showing this thing. In this manner, in closeness of a readiness set we should simply process the unforeseen probabilities since the structure is unique. At the point when the framework is estimated, it is possible to describe any new thing giving its characteristics using the Baye's standard. The Naïve Bayes request technique is gotten in this examination essentially for its ability to manage missing features, which occurs for a part of the neuropsychological assessments. A Naïve Bayes classifier is a fundamental probabilistic classifier reliant on the utilisation of Bayes' speculation (portrayed numerically underneath) with the assumption of probabilistic self-rule between each pair of features; essentially this is rarely legitimate, as explicit features can be connected, yet Naïve Bayes classifiers display astoundingly generous execution on features which are not cautiously self-sufficient (N. Amor, 2004). Figure 26 shows the code which is used to create Gaussian model.

Gradient Boosting

The boosting method generates base models sequentially. Prediction accuracy is improved through developing multiple models in sequence by putting emphasis on these training cases that are difficult to estimate. In the boosting process, examples that are difficult to estimate using the previous base models appear more often in the training data than the ones that are correctly estimated. Each additional base model is aimed to correct the mistakes made by its previous base models. A weak leaner is an algorithm that performs only slightly better than random guessing; a strong base model is a more accurate prediction or classification algorithm that is arbitrarily well correlated with the problem. The answer to this question is very important. It is often easier to estimate a weak model compared with a strong model. Schapire (Kearns, 1988) proves that the answer is positive by applying boosting algorithms to combine many weak models into a single and high accurate model.

Friedman proposed a modification to the gradient boosting method by using a regression tree of fixed size as the base model. The modified version improves the quality of the model. In this study, this modified version of the GBM model is used for travel time prediction. Assuming that the number of leaves for each tree is J. Each tree partitions the input space into J disjoint regions R1m, R2m, ..., Rjm and predicts a constant value bjm for region Rjm. The regression tree can be formally expressed as:

$$g_m(x) = \sum_{j=1}^{J} b_{jm} I(x \in R_{jm})$$

MLP

Multi-Layer Perceptrons (MLP) are completely associated feedforward nets with at least one layers of hubs between the info and the yield hubs. Each layer is made out of at least one counterfeit neurons in parallel. A neuron, has N weighted sources of info and a solitary yield. A neuron consolidates these weighted contributions by shaping their entirety and, with reference to a limit worth and actuation work, it will decide its yield. Figure 6 shows the idea of this algorithm.

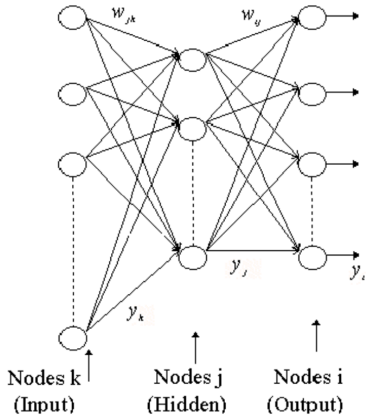

Figure 6. MLP algorithm

The net is prepared by at first choosing little arbitrary loads and inside limits, and showing all preparation information more than once. Loads are balanced after each preliminary utilizing data determining the right class until loads unite and the cost capacity is diminished to a worthy worth. The by and large great execution found for the back proliferation calculation is to some degree astounding thinking about that it is a slope drop system that may locate a neighbourhood least in the cost capacity rather than the ideal worldwide least (Riedmiller, 1994).

Experiments Design and Implementation

The experiments of the project is composed of several main steps which include data understanding, data preparation, data transformation, model selection. After that, I also conducted grid search to find the best parameters for selected models and then perform iterations to refine the result. Figure 7 shows the workflow of the experiment.

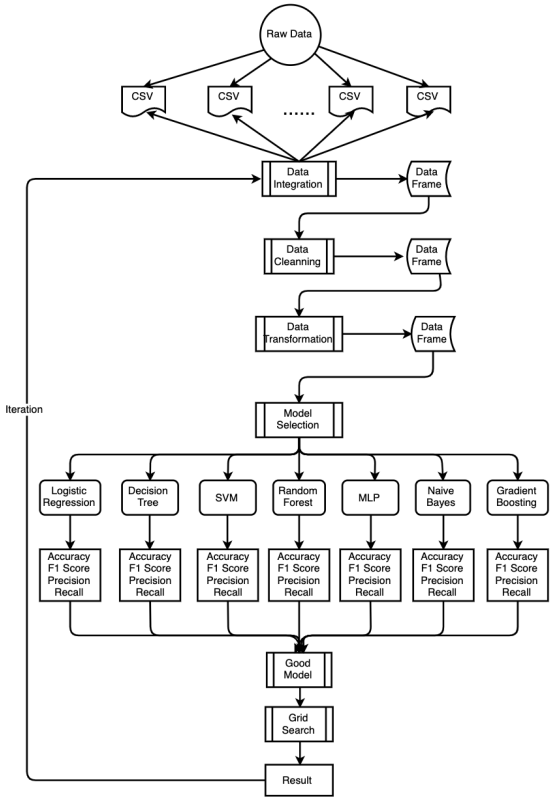

Figure 7. Design of Experiment

Data Understanding

In this phase, I will continue to take a closer look at the data available for mining. It is critical to have a good understanding of data and avoid protential problems during data preparation. In this section, I will access the data and explore it using tables and graphics. This makes it possible for me to determine the quality of the data and describe the results of these steps in the project documentation.

Collect initial data

Existing Data. The data is available from UCI Machine Learning Repository. It contains part of the records of Census database in 1994 and is extracted by certain rules. The dataset contains all of the information of adult including income and other information like education and occupation. The format of the dataset is csv,

and size of the dataset is also able to meet the requirement of this project. There are some attributes will be considered to be useless in the following steps..

Describe the data

The descriptions of dataset in this project are focusing on the quantity and quality of the data. Listed below are some key characteristics of the dataset.

Amount of data. For most techniques of models, there are trade-offs correspond with data size. Large volume of data sets can generate models which are more accurate, however they can also increase the length of the processing time. In this project, there are totally 15 fields in the table with 48842 records in the dataset.

Value types. The describe function of pandas package is used to describe the continuous variables of dataset. According to the result from Figure1, there are 6 continuous variables type which are age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week. The others are all categorical variables. According to the result, the average age is about 38.

Explore the data

To address the data mining goal, I explored the dataset using charts and visualisation tools including plotting the density and counting of raw data and applying basic statistic method. This can also help to formulate the hypotheses of the data and shape the data transformation tasks which will take place during the step of data preparation.

Figure 8 shows the density graph of age and hours-per-week. The graph shows that the majority of age adult in dataset is between 18 to 60 years old. For adults of above 60 years old, the data may be able to be removed since they are almost at retire age. According to the graph, the working hours of adult is mostly about 40 hours, and this is reasonable because the average working hour is eight hour per day with five official working dat per week. More analysis will be conducted during the data preparation step.

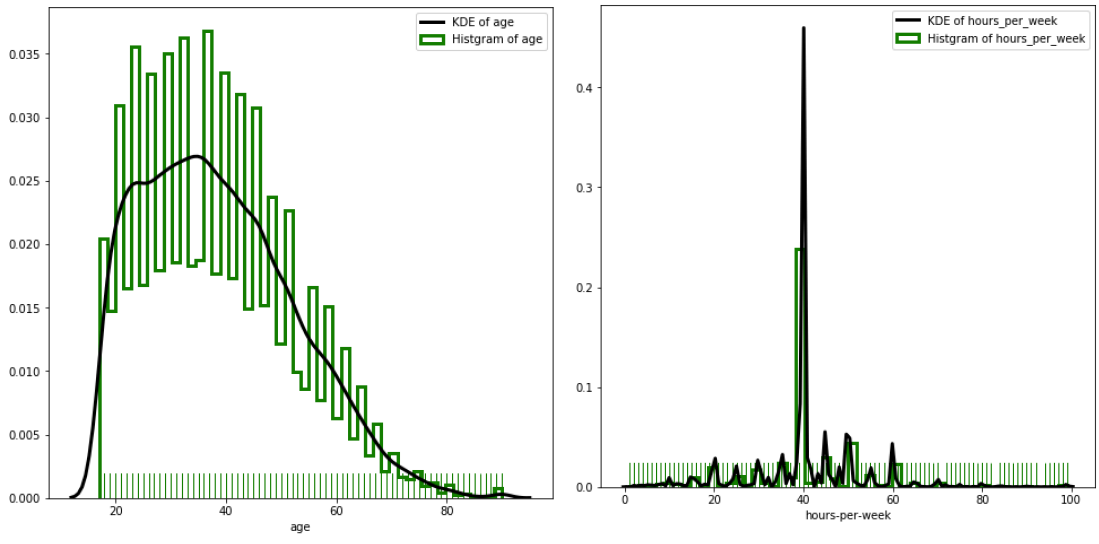

Figure 8. Density plot of age and hours-per-week

Figure 9 shows the density graph of capital_gain and capital_loss. The graph shows that the majority of adult in dataset have zero values of capital_gain and capital_loss, so we can drop these two columns in the dataset.

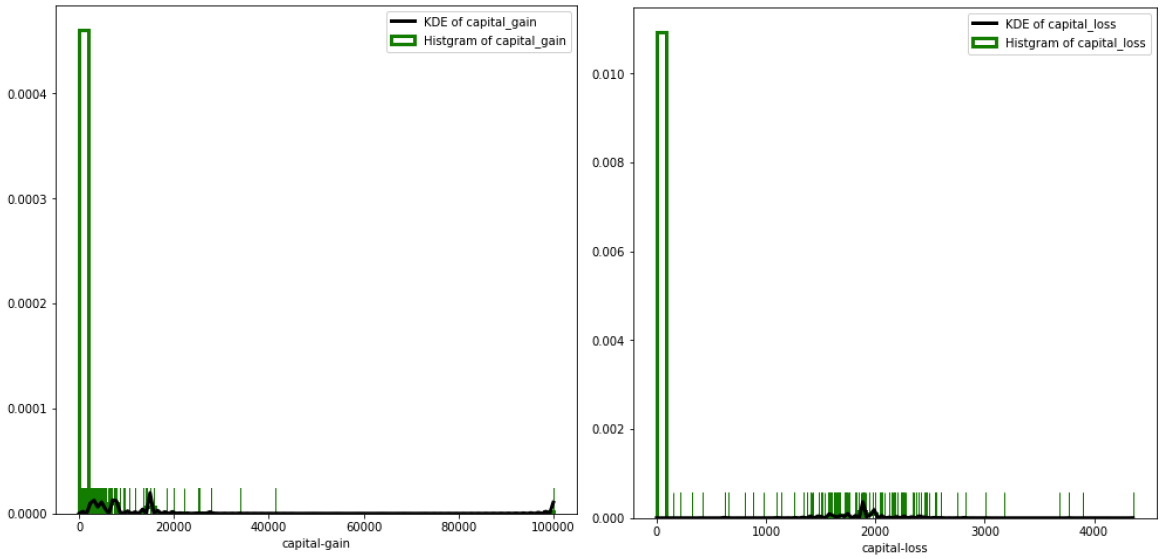

Figure 9. Density plot of capital_gain and capital_loss

Figure 10 shows the distribution of education and marital-status. According to the above graph, there are three categories of education which are of higher percentage than others including Bachelor,HS-grad and Some-college. The percentage of them is 16.4%, 32.2% and 22.3% respectively. Other categories of education are of lower percentage which are below 5%. The graph also shows that there are 45.8% of adults are of Married-civ-spouse category. About 33% of the adults who have never been married and 13.6% of them are divorced. Other categories of them are of lower percentage which is less than 5%.

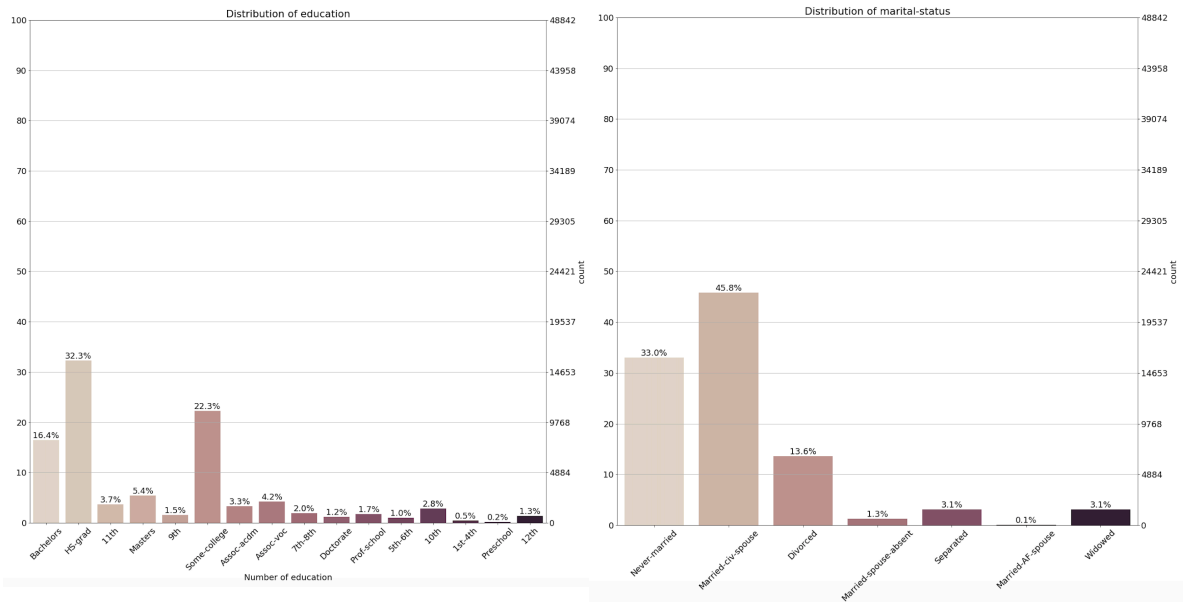

Figure 10. Distribution of education and marital-status

Figure 11 shows the distribution of race and native-country. According to the above graph, the percentage of white people is much higher than other categories which is about 85.5%. The percentage of black people is about 9.6%. The percentage of people in other race is all lower than 5%. The graph also shows that the majority of adults are from United States, and the percentage reached to about 89.7%. This also indicated that the native-country maybe not so important in this project. It's possible to consider to drop this column in the following steps.

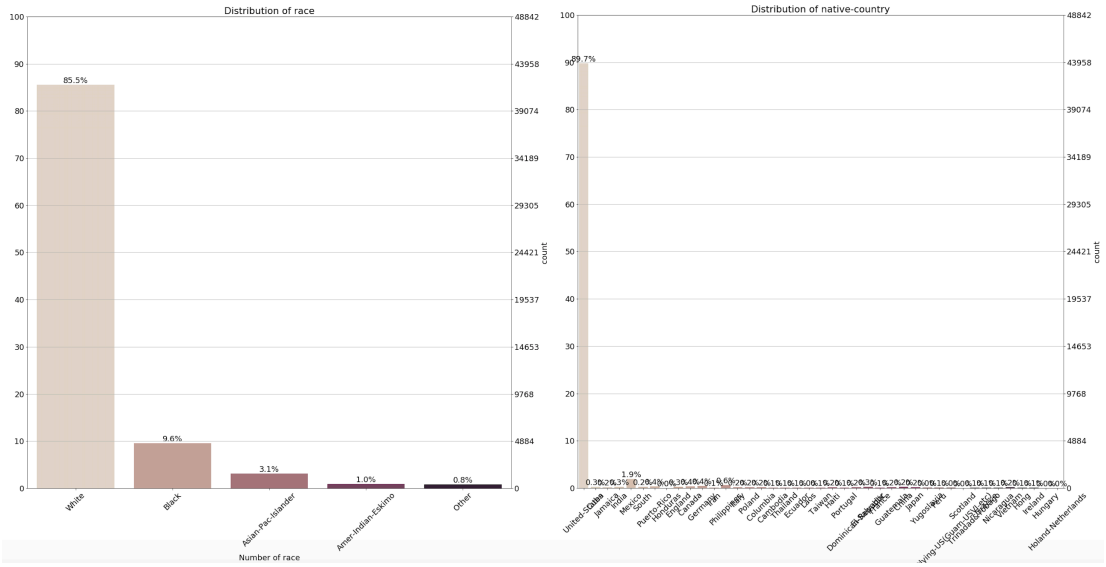

Figure 11. Distribution of race and native-country

Figure 12 shows the distribution of work-class and occupation. According to the above graph, the majority of adults is doing private business and the percentage of it is about 69.4%. For other types of work class like state-gov, self-emp-not-inc and so on, the percentage of them are much lower. The graph also shows the distribution of occupation which looks more balanced. The percentages of most occupations are around 10%. There are some categories like protective-serv and priv-house-serv, of which the percentage is almost 0.

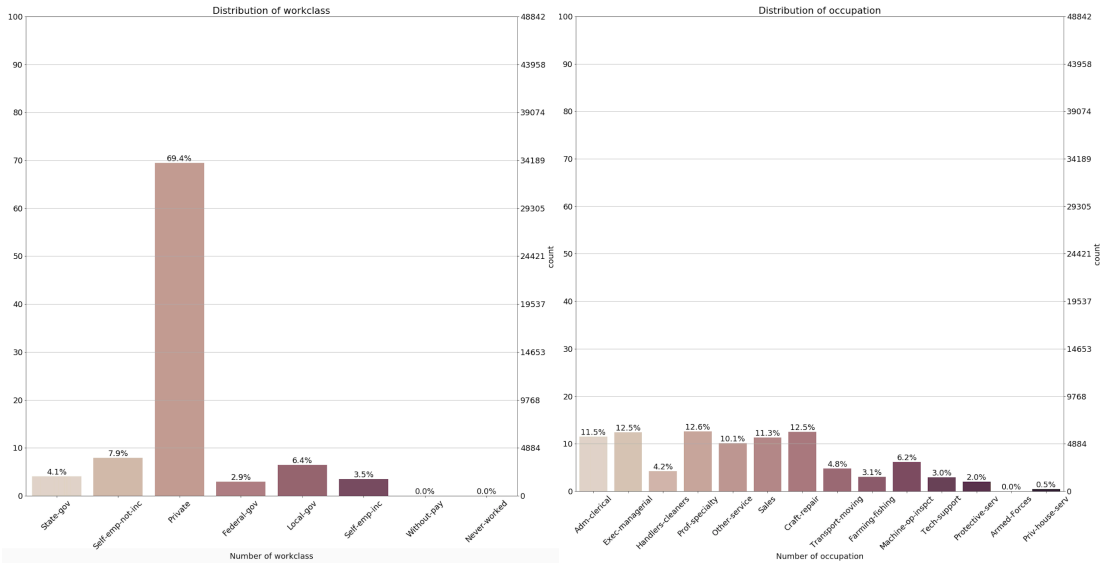

Figure 12. Distribution of workclass and occuxpation

Figure 13 shows the distribution of relationship in family and the income of adult. The graph indicates that there are about 40 percentage of adults are acting as husband and there are about 25 percentage of adults are not in a family. The percentage of adults who owns a child is about 15.5%. There are about 10 percentage of adults who are not married. The graph of income indicates that there are about 76 percentage of adults of which the income is less than 50K per year, and the rest of them are larger than 50K per year.More analysis will be conducted during the data preparation step.

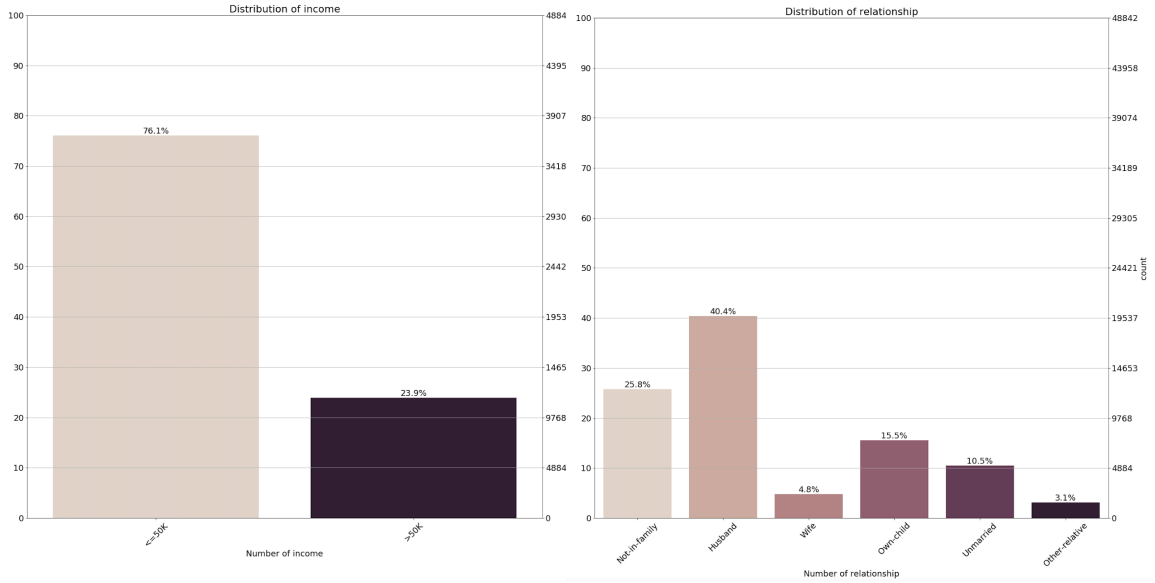

Figure 13. Distribution of relationship and income

Besides the plot of density and distribution, I also conducted exploration about the relationship of income and other factors including age, hours-per-week and education-num and generated boxplot. Figure 10 shows the code of boxplot.Figure 14 shows the result of boxplot.

According to the result of box plot, the average age of adults who have income of larger than 50K is higher than the adults who have income of less than 50K. The adults of higher education level are more likely to get income of larger than 50K according to the result of the boxplot. The situation is almost the same with hours-per-week which means the adults who have income larger than 50K also have longer working hours per week.

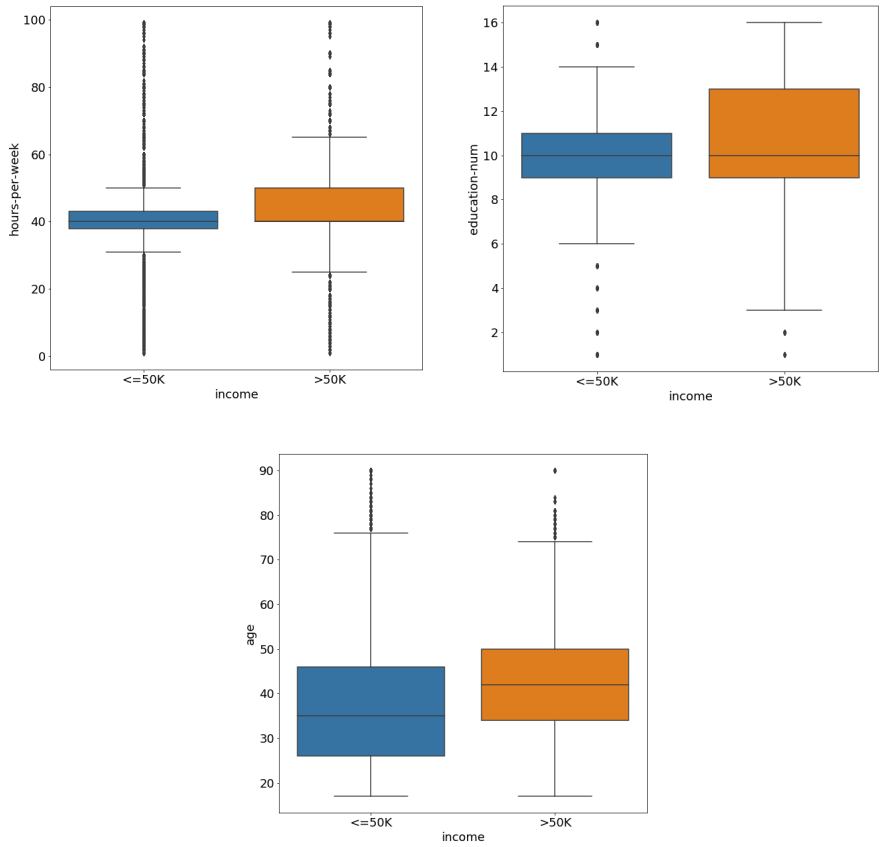

Figure 14. Distribution of relationship and income

Verify the data quality

After basic exploration of dataset, it's also critical to ensure the quality of dataset. Before starting the modelling, what is more important it to ensure that the dataset does not contain errors like missing values, data errors and measurement errors. To avoid potential pitfalls, I conducted a quality analysis of available data throughly before modelling. Figure 14 shows the code which is used to conduct the testing.

- ***Missing data.*** The attribute 'workclass','occupation' and 'native-country' has missing values with percentage of 6%, 6% and 2%.
- ***Data errors.*** I conducted the testing of checking if there is character like "?" In the dataset, and according to the testing of data errors, there is no such errors in the dataset.
- ***Measurement errors.*** In this project, zero values are taken as measurement errors. There are 92 percent of records have zero values for attribute 'capital-gain' and 95 percent of that for attribute 'capital-loss'.

Data preparation

Data preparation is one of the most important and often time-consuming aspects of data mining. In fact, this step takes about 60% of the time and effort of this project. According to the goals of the project, I conducted data preparation which involves several different tasks, including merging data, selecting a sample subset of data, removing blank or missing values and splitting the dataset into training and testing datasets.

Select the data

Selecting items. The study will be limited to the approximately 42882 adults, and I selected the records which the age of adult is between 18 to 60 and this will also reduce the data of outliers.

Selecting attributes. The basic line to select the attributes in this project is that the percentage of zero values can not be greater than 50 percent. According to the result of data quality, attributes like 'capital-gain' and 'capital-loss' have above 90 percent of zero values, so these two attributes will be dropped during the following steps. In the first iteration, all the attributes will be kept to see the result. Also, I conducted the testing of importance for features using different ways and results will be compared in the other iterations for this project. Details of this will be illustrated in Chapter 8.5.

Clean the data

According to the result of data validation of quality, there are missing data and measurement errors in the dataset, and the invalid data need to be cleaned before modelling.

Missing data. The attribute 'workclass','occupation' and 'native-country' has missing values with percentage of 6%, 6% and 2%.Data with missing values have been deleted from the dataset.

Measurement errors. There are 92 percent of records have zero values for attribute 'capital-gain' and 95 percent of that for attribute 'capital-loss'. So the two columns are dropped from datset.

Format the data as required

As a last step before modelling, it is useful to check whether certain procedures require a specific arrangement or request to the information. For instance, it isn't unprecedented that a succession calculation requires the information to be pre-sorted before running the model. For the models I used in this project, no particular data format or order is required, however, there are several categorial variables in the dataset, to proceed with the modelling, I transformed the categorial variables to be numeric by one hot encoding.

Data transformation

Reduce the data

In the dataset, the column “fnlwgt” is considered useless in this project, so I dropped this column. Also feature importance of all the attributes are generated using altirigom of decision tree. Figure 15 shows the feature importance.

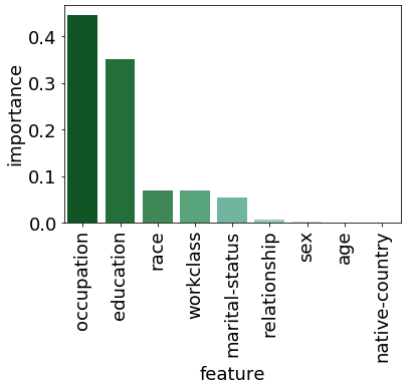

Figure 15. Feature Importance

The result shows that sex, age, and native-country has much less importance score than other attributes, they are supposed to be dropped. In the first iteration of this project, I kept all above attributes, and then dropped features of less importance to see the difference. Figure 18 shows the code which is used in this step.

Project the data

After selecting the importance feature, new dataset is generated. Besides transforming the categorical variables in to labels, I conducted one hot encoding to transform all the categorial data into a format which is easier for computer to understand.

Data-mining method selection

The most appropriate mode is determined based on the following considerations:

- ***The data types available for mining.*** The data types of input for this project are numeric and categorical and the categorical variables will be transformed to labels before the modelling.The data type of target is categorical variable.
- ***Data mining goals.*** There are two main steps for the objectives of this project. One is to create new model to train the dataset and distinguish it into different groups. After this, the model is used to predict if the income of an adult will be higher than 50K. The objectives match the method of classification among different data mining methods.

- ***Specific modelling requirements.*** No particular size of data is required for the model but it does require variables to be taken as target. Easily presentable results are needed after building the model.

According to above discussion, classification is the appropriate method of data mining for this project.

Data mining and Result

The experiment conducted data mining using different models including SVM, Decision Tree, Random Forest, Logistic Regression, MLP, Naive Bayes and Gradient Boosting for classification and evaluate the result with evaluation metrics.

According to the result, logistic regression and gradient boosting are selected as the model for the project because of the higher level of accuracy and f1 score. Table 1 shows the result of data mining.

Table 1. Result of data mining

	Accuracy	F1 Score	Precision	Recall
Logistic Regression	0.77759	0,743661	0,751	0,77759
Decision Tree	0.774651	0,749166	0,748753	0,774651
Random Forest	0,760838	0,67383	0,736414	0,760838
SVM	0,755033	0,649646	0,570075	0,755033
Naive Bayes	0,757678	0,757182	0,7567	0,757678
MLP	0,772193	0,731605	0,741697	0,772193
Gradient Boosting	0.776194	0,774234	0,749189	0,77568

Grid Search

To refine the result of models, grid search is applied to Logistic Regression and Gradient Boosting to find the best parameters and result. Table 2 shows the result of grid search.

Table 2. Result of Grid Search

	Gradient Boosting	Logistic Regression
Accuracy	0.776194	0,774234
F1 Score	0.75165	0,741764
Precision	0,751144	0,747126
Recall	0.776194	0,774234

Interpretation and Results

According to the result of correlation matrix, attributes like work-class, education, occupation, race, sex have affect on the value of income. This is almost the same with the result of feature importance got from models. It's also the foundation for further research about conducting more iterations to improve the model and find patterns like the relationship between income to these factors.

To analyse the mined pattern, more graphs of relations between multiple factors are generated to help on interpretation.

Figure 36 shows the relationship between income, marital-status and education-num.

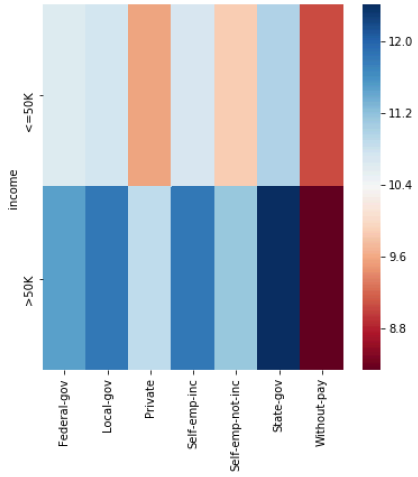

Figure 36. Heat Map of Income, Marital-status, Education

Figure 37 shows the relationship between income, education-num and workclass.

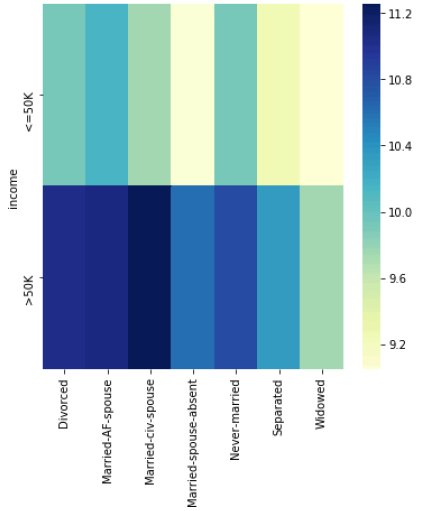

Figure 37. Heat Map of Income, Education-num, Workclass

Figure 38 displays the pattern between income, occupation and education-num.

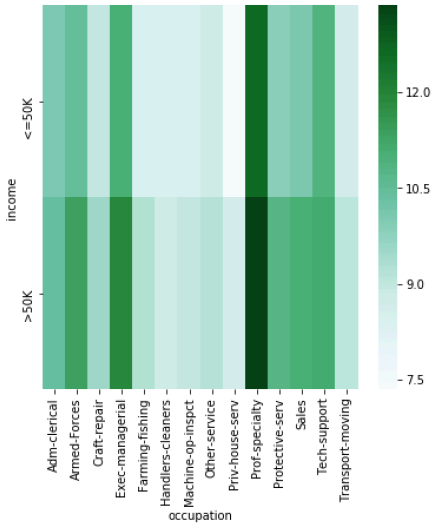

Figure 38. Heat Map of Income, Education-num, Workclass

According to the result of analyse, people with higher education level and married would have more chance to get income which is larger than 50K. At the same time, people who work for government with higher level of education would also have high salary. Age is not an important factor to income because people with different age would have various level of income.

The model which has the highest f1-score is gradient boosting and logistic regression in the experiment of the first iteration and the f1-score 0.77. Also more experiments and analyse will be conducted in the other iterations.

Conclusion

According to the above steps, there are a couple of elements which are of bigger measure of centrality inlacing model, relationship, occupation, conjugal status, instruction. People who have increasingly raised measure of preparing and work for government would have continuously chance to get pay which is greater than 50K. Both logistic regression and gradient boosting can be utilised to go before with this investigation and the accuracy of the model is about 0.78 which is higher than that of various models.

As per the result of the project, I proposed the action which would be necessary to conduct. To apply the information and convey the execution, government should make more effort to raise the average level of education for citizens especially for urban areas. Government should also support private business to raise the average salary for workers so that the salary would be balanced between normal employers and those who works for government. Expecting that they acknowledge the consequences of the investigation, instruction ought to be of higher need for the commitment of urban areas.

Further Work

To screen and keep up the execution, the qualities like relationship, occupation, conjugal status ought to be followed later on. The code of the undertaking was executed in python bundles which can be utilized to get the legitimacy and precision of the model. A model of f1 score under 0.7 will be considered as terminated. After the model is lapsed, the suit of the code can be utilized to run the models and get the new model which has most astounding precision.

To improve the arrangement later on, more emphasess can be utilized to execute the venture consequently when the dataset is refreshed. Progressively mind boggling and solid representation can be applied to locate the more profound patten between various variables.

References

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer Relationship management: A literature review and classification. Expert Systems with Applications, 36(2, Part 2), 2592-2602.

Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang. PEBL: positive example based learning for web page classification using SVM. In KDD, pages 239–248. ACM, 2002.

Loh (2011), Classification and regression trees, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

C L Blake, C J Merz. UCI repository of machine learning databases University of California, Irvine, Department of Information and Computer Sciences. 1998

Jakovljevic, N., et al. "Comparison of linear discriminant analysis approaches in automatic speech recognition." *Elektronika ir Elektrotechnika*, vol. 19, no. 7, 2013, p. 76+. *Gale Academic Onefile*. Accessed 13 Sept. 2019.

Shi, H., and Liu, Y. (2011). Naïve Bayes vs. support vector machine: resilience to missing data, in Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7003 LNAI(PART 2) (Taiyuan), 680–687.

J. Wang, S. J. Redmond, M. Bertoux, J. R. Hodges, and M. Hornberger, “A comparison of magnetic resonance imaging and neuropsychological examination in the diagnostic distinction of Alzheimer’s disease and behavioral variant frontotemporal dementia,” Frontiers in Aging Neuroscience, vol. 8, article 119, 2016.

O. Caelen, “A Bayesian interpretation of the confusion matrix”, Annals of Mathematics and Artificial Intelligence, 81(3–4), 429–450, 2017.

Riedmiller, M.A. (1994). Advanced supervised learning in multi-layer perceptrons — From backpropagation to adaptive learning algorithms.

Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, “Efficient KNN classification algorithm for big data,” Neurocomputing, vol. 195, pp. 143–148, Jun. 2016.

M. E. Mavroforakis and S. Theodoridis, “A geometric approach to Support Vector Machine (SVM) classification,” IEEE Trans. Neural Netw., vol. 17, no. 3, pp. 671–682, May 2006.

Safavian, S. R., and Landgrebe, D., 1991, A survey of decision tree classifier methodology. IEEE T ransactions on Systems, Man, and Cybernetics, 21, 660–674.

Liaw A, Wiener M (2002). “Classification and Regression by randomForest.” R News, 2(3), 18–22.

N. Amor, S. Benferhat, and Z. Elouedi. Naive bayes vs decision trees in intrusion detection systems. In Proceedings of the 2004 ACM symposium on Applied computing, pages 420–424. ACM, 2004.

Balakrishnama, S., Ganapathiraju “Linear Discriminant Analysis – A Brief Tutorial” Institute for Signal and Information Processing, 1998.

King, Gary, and Langche Zeng. 2001. “Logistic Regression in Rare Events Data.” Political Analysis, in press.

Madhu Sanjeevi. (2017). Different types of Machine learning and their types. Retrieved from www.medium.com/deep-math-machine-learning-ai/different-types-of-machine-learning-and-their-types-34760b9128a2

Statistical Solution. (2019). What is Logistic Regression. Retrieved from www.statisticssolutions.com/what-is-logistic-regression/

Medcalc. (2019). Logistic regression. Retrieved from www.medcalc.org/manual/logistic_regression.php

R. Berwick. (2019). An Idiot’s guide to Support vector machines (SVMs). Retrieved from web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf

Rajesh S. Brid . (2018). Introduction to Decision Trees. Retrieved from www.medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

Yellowbrick. (2019). Classification Report. Retrieved from www.scikit-yb.org/en/latest/api/classifier/classification_report.html

GeeksforGeeks. (2019). Confusion Matrix in Machine Learning. Retrieved from www.geeksforgeeks.org/confusion-matrix-machine-learning/

Wiki. (2019). Log Loss. Retrieved from www.wiki.fast.ai/index.php/Log_Loss

S. Chakraborti, “A Comparative Study of Performances of Various Classification Algorithms for Predicting Salary Classes of Employees,” vol. 5, no. 2, pp. 1964–1972, 2014.

Khongchai, P., & Songmuang, P. (2016). Random Forest for Salary Prediction System to Improve Students? Motivation. In 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 637?642). IEEE. https://doi.org/10.1109/SITIS.2016.106

Viroonluecha, P., & Kaewkiriya, T. (2018). Salary Predictor System for Thailand Labour Workforce using Deep Learning. *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*, 473-478.

I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright. (See: https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data.