CONTENT

Title	Topic
• ABSTRACT	1
• INTRODUCTION	2
• OBJECTIVE	3
• METHODOLOGY	4
• EXPERIMENTATION	5
• RESULT and DISCUSSION	6
• CONCLUSION	7
• FUTURE SCOPE	8
• REFERENCE	9

ABSTRACT

Subsidy Income is a company who delivers subsidies to individuals based on their income. But now a days, it's very much difficult to know the actual income of every individual. So accurate income data is one of the hardest piece of data to obtain across the world. Subsidy Inc. has obtained a large data set of authenticated data on individual income, demographic parameters, and a few financial parameters. But collection of fully accurate data may not be possible at all. Due to the non-availability of accurate data on financial parameters, incompleteness in regard to the coverage is embedded in such the dataset.

Subsidy Income company wishes to develop an income classification-system for individuals using some statistical methods to enhance the accuracy-level of the data and to create a good classification using some high performance algorithms, like, Logistic Regression algorithm ,K-Nearest Neighbours Classifier algorithm and Decision Tree algorithm using python software. The above Algorithms have been used to develop classifier systems to measure the number of misclassifications which leads to the criterion of measuring accuracy embedded in the algorithms with respect to a dataset.

INTRODUCTION

Subsidy Income is a company who delivers subsidy to various customers according to their income level. A subsidy is basically a quantity of money given directly to companies, organizations or individuals by the Government, i.e. ,the taxpayer. Subsidies aim to encourage productions, boost exports and prevent a business from collapsing. Here Subsidy Income delivers subsidies to customers according to their need for business, productions or any other purpose. But the amount of subsidies are different according to the income level of each individual. So a person who has higher income level, may not get subsidy but a person whose income level is not good, get the subsidy for any kind of startup or any other purpose for income enhancement. In the modern world, collection of the correct information about the income of an individual is not easy. Many people don't want to disclose the correct information about their income.

So it is very much necessary to classify the income level for each and every individual with higher accuracy for the betterment of business planning and strategy. Subsidy Income company collects an worldwide data according to the income of the individual, some demographic features and financial parameters. To classify the individual's income with high accuracy, some statistical techniques are applied. Here some machine learning algorithms are used for the betterment of classification. The dataset which is used to solve this classification problem has 31978 rows and 13 columns. Each and every column is known as a parameter. Here are some categorical columns having different categories. Using all the parameters, the classification of income level has been done. 'SalStat' is a column on which the classification will be done i.e. this is a dependent variable which is categorical in nature. This column has two categories- 1. more than 50,000($ per annum) and 2. less than or equal to 50,000($ per annum).

There are some other parameters such as-

```
age             31978 non-null int64
JobType         31978 non-null object
EdType          31978 non-null object
maritalstatus   31978 non-null object
occupation      31978 non-null object
relationship    31978 non-null object
race            31978 non-null object
gender          31978 non-null object
capitalgain     31978 non-null int64
capitalloss     31978 non-null int64
hoursperweek    31978 non-null int64
nativecountry   31978 non-null object
SalStat         31978 non-null object
dtypes: int64(4), object(9)
```

Here ‘age’- age of individuals
‘JobType’- type of their works
‘EdType’- individual’s educational qualification
‘maritalstatus’- individuals are married or not
‘occupation’- occupation of individuals
‘relationship’- wheather an individual has family, children etc.
‘race’- race of the individual (skin colour of individual)
‘gender’- gender of individual
‘capitalgain’- capital gain of individual(Profit while selling a property)
‘capitalloss’- capital loss of individual(Loss while selling a property)
‘hoursperweek’- working hours per week of an individual
‘nativecountry’- country every individual belongs to
‘SalStat’- income of each and every individual

After using some data cleaning operations and visualizations, the independent variable ‘SalStat’ is classified using appropriate dependent variables using some machine learning algorithms like Logistic Regression, Kneighbors Classifier and Decision Tree Classifier. Accuracies for different algorithms are recorded. The algorithm having highest accuracy and minimum number of misclassifications is likely to use to solve the classification problem.

OBJECTIVE

The objective of this project is:-

1. Simplify the data system by reducing the number of variables to be studied, without sacrificing too much accuracy.
2. Reduce the number of misclassifications in the data.
3. Find the best algorithm among some algorithms according to the accuracy and number of misclassifications of this classification problem to find out accurate income of a person.

METHODOLOGY

In this classification problem, some statistical methods and various algorithms are applied to reach at the objective. Having the dataset, some data cleaning operations are applied to clean some preliminary defaults like wrong entries, missing values etc. .After the data cleaning operation, the visualization part comes into play. Here the relation between the dependent variable 'SalStat' and other independent variables are checked using some appropriate graphical methods. The relation between two categorical variable are checked using crosstabulation method i.e. using contingency tables. Using these methods, one can be assure that which of the independent variables are relevant to explain the variability in the dependent variable('SalStat') i.e. the variables which are irrelevant to classify the dependent variable 'SalStat', are removed. Then dummy variable columns are created in place of the categorical variables in the dataset. After creating dummy variables, the whole dataset is divided in two parts- 1. dataset having all relevant independent variables, 2. a column of dependent variable('SalStat'). The size of the matrix of the independent variables is (30,162 ,94) and the size of the independent variable is (30,162 , 1).

Then comes some machine learning techniques to solve the rest and one of the most important parts in this project. Using the supervised learning method, the dataset of the both independent and dependent variables are split into training set(70% of the whole dataset) and testing set(30% of the whole dataset) randomly. Then using logistic regression algorithm, the logistic model is built using the training dataset for both independent and dependent variable. Then the validation (accuracy) of this classification model is checked using the testing set of both independent and dependent variable. The number of misclassifications are also recorded. The predicted values corresponding the model are listed as an output. The column of the output values has only two categories which are 0 and 1. 1 means the individual has income level more than 50,000$ and 0 means the individual has income less than or equal to 50,000$. To see wheather the accuracy can be improved or not, other 2 algorithms are applied for using similar way. Those algorithms are K-Nearest Neighbors classifier and Decision Tree classifier.

While using the K-NN algorihm(K-Nearest Neighbour) and Decision Tree algorithm, the parameters of these two algorithms are set in iterative method to reach the optimum solution i.e. the maximum accuracy and minimum number of misclassifications.

Among these 3 algorithms of supervised learning, the best one can be selected according to the maximum accuracy and minimum number of misclassifications. That algorithm is likely to solve the classification problem with much higher accuracy.

The brief descriptions of some special techniques which are applied in these problem, are given below:

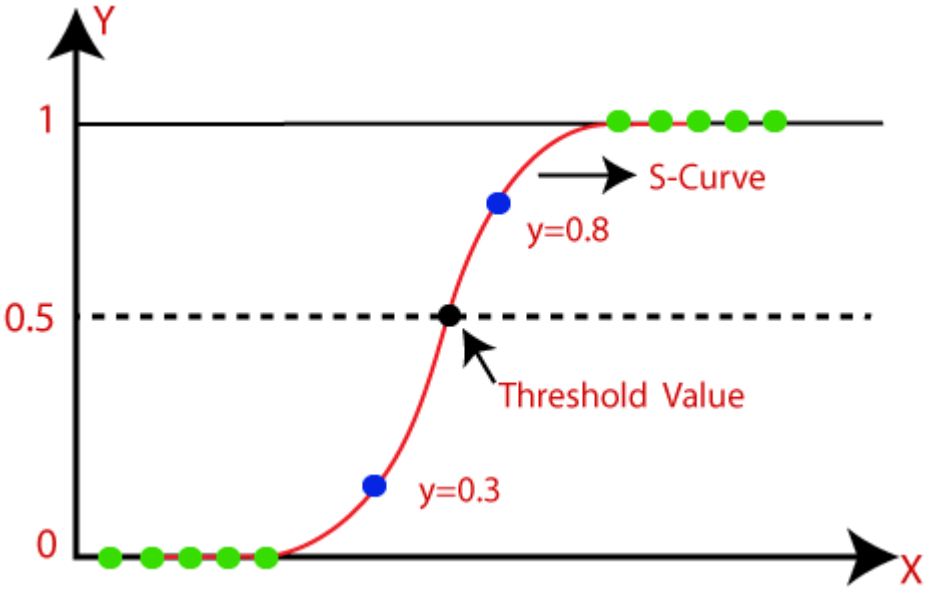
Missing value treatment: In this classification problem, there are some missing data present in the dataset which decrease the accuracy of the model. The necessary steps have been taken to solve this issue. The missing values in the categorical columns are replaced by the mode i.e. the maximum occurace of a particular category. Mean and median are recorded for all numerical columns. The missing values of the column whose mean and median are not so much far from each other, are replaced by the mean. Otherwise, the missing values for numerical columns are replaced by median because it is the best measure of central tendency.

Training set and testing set: The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make prediction on the data not used to train the model. It is a fast easy procedure to perform to check the performance of the model quickly. In this algorithm, the whole dataset is divided in 2 parts- training set and testing set. Usually the ratio of this split is 80%:20% or67%:33% or 70%:30%. This split is happened randomly on the dataset.

Logistic Regression algorithm: Logistic regression is classification algorithm used to assign observations to discrete set of classes. Logistic regression transforms its output using the logistic sigmoid function to return a probability value. Sigmoid function is basically a function who maps any real value into another value between 0 and 1. In machine learning, sigmoid function is used to map predictions to probabilities.

Assumptions for Logistic Regression:

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity

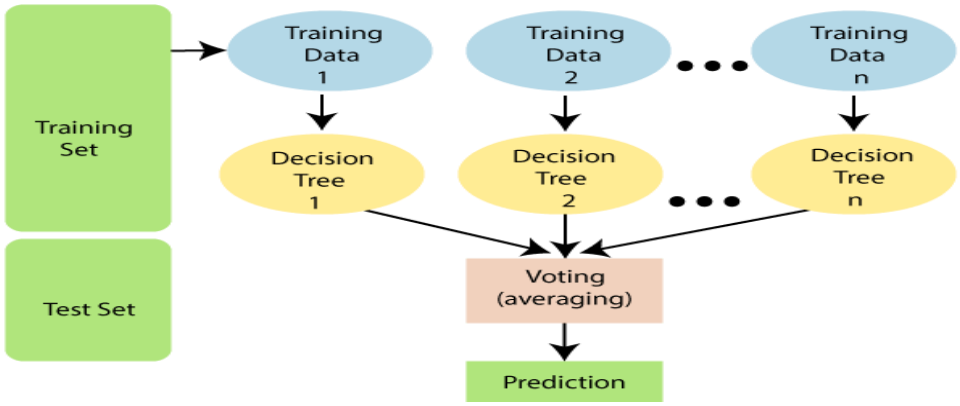


K-NN algorithm: K-Nearest Neighbour algorithm is one of the simplest supervised machine learning algorithms. K-NN algorithm assumes the similarity between the new case/data and the available cases/data and put the new case into the category that is most similar to the available categories. It is a non-parametric algorithm which means it does not make any assumption on the underlying data. It is also called a lazy learning algorithm because instead of learning from the training set, it stores the dataset and at the time of classification, it starts an action on the dataset.

Decision Tree Algorithm: Decision Tree algorithm belongs to the family of supervised learning algorithms. It is used to solve both classification and regression problems. The goal of this algorithm is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from the prior data(training set). In this algorithm, for predicting a class label for a record we start from the root node of the tree. The values of the root attribute are compared with the record's attribute. On the basis of comparison, the branch is followed corresponding to that value and jump to the next node.

Random forest : Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance .According to what its name implies, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that

dataset." Instead than depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. **The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**
The below diagram explains the working of the Random Forest algorithm:



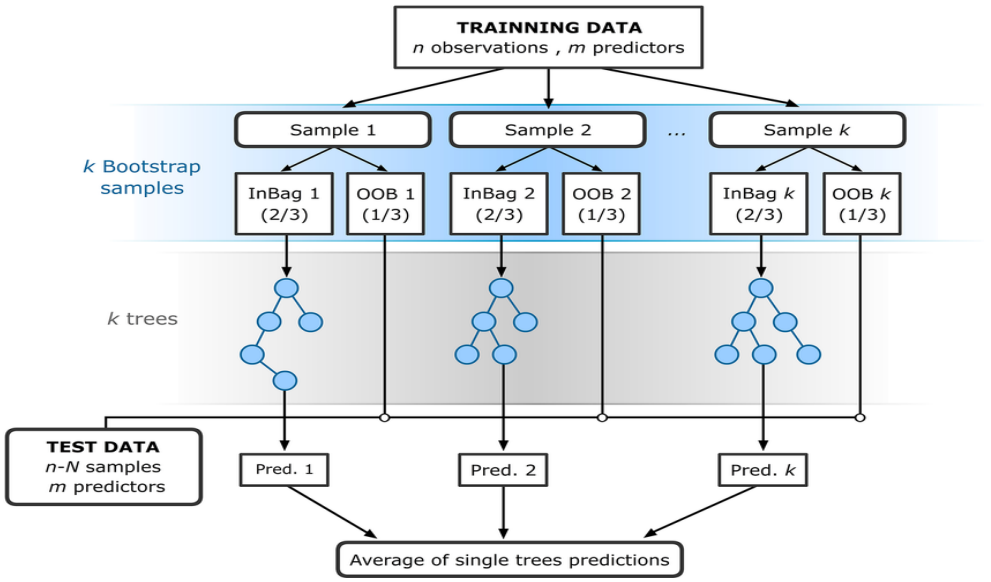
Assumptions for Random Forest :

Some decision trees may predict the correct output, while others may not, because the random forest combines numerous trees to forecast the class of the dataset. But when all the trees are combined, they forecast the right result. Consequently, the following two presumptions for an improved Random forest classifier:

- For the dataset's feature variable to predict true outcomes rather than a speculated result, there should be some actual values in the dataset.
- Each tree's predictions must have extremely low correlations.

How does Random Forest algorithm work?

First, N decision trees are combined to generate the random forest, and then predictions are made for each tree that was produced in the first phase.
The stages and graphic below can be used to demonstrate the working process:
Step 1: Pick K data points at random from the training set.
Step 2: Create the decision trees linked to the subsets of data that have been chosen.
Step 3: Select N for the size of the decision trees you wish to construct.
Repeat steps 1 and 2 in step 4.
Step 5: Assign new data points to the category that receives the majority of votes by looking up each decision tree's predictions for the new data points.



Data Analysis and Visualisation

After doing some data cleaning operations on the dataset in python, some special visualization techniques are done to solve the problem more clearly and easily. Here some experiments are done on the numerical as well as categorical columns to see the relationship between the independent variables/columns and the dependent variable/column('SalStat').

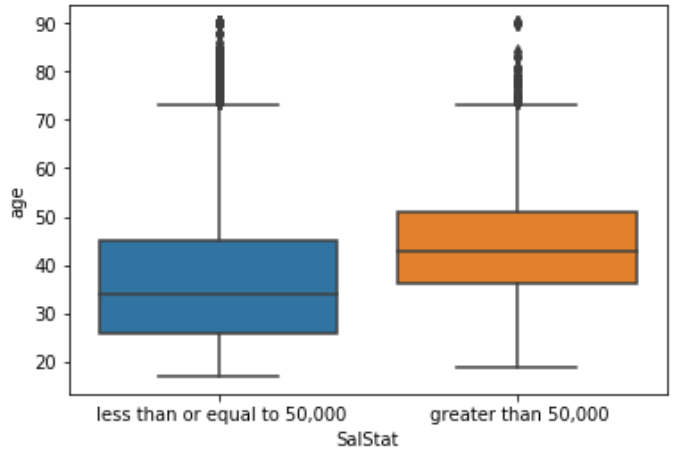
Some important details about those experiments are given below:

Gender and SalStat:

SalStat	greater than 50,000	less than or equal to 50,000
gender		
Female	0.113678	0.886322
Male	0.313837	0.686163
All	0.248922	0.751078

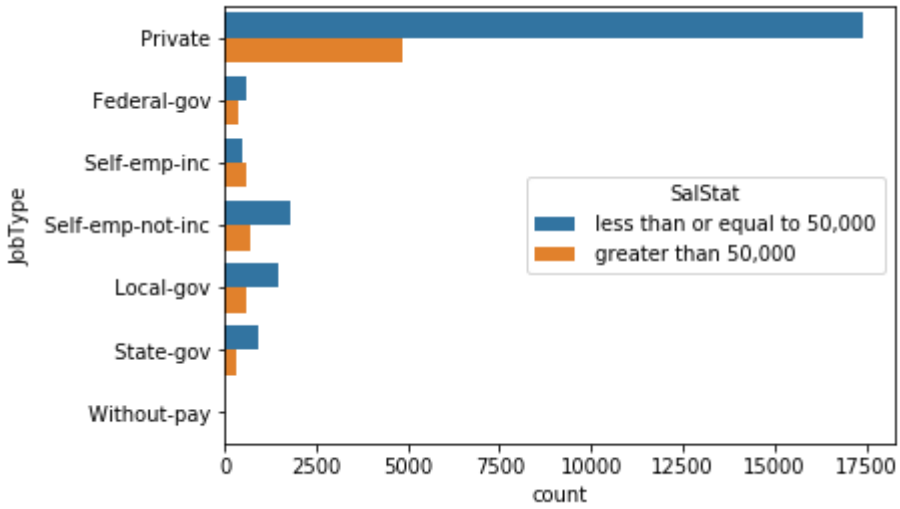
According to the dataset, this table shows that 88.63% of female category has income-level less than or equal to 50,000$ and 68.61% of male category has income-level less than or equal to 50,000$. 31.38% of male category has income-level more than 50,000$.

Age and SalStat:



People with age 35-50 income level greater than 50,000$ and
people with age 25-35 income level less than or equal to 50,000$ according to the dataset.

JobType and SalStat:

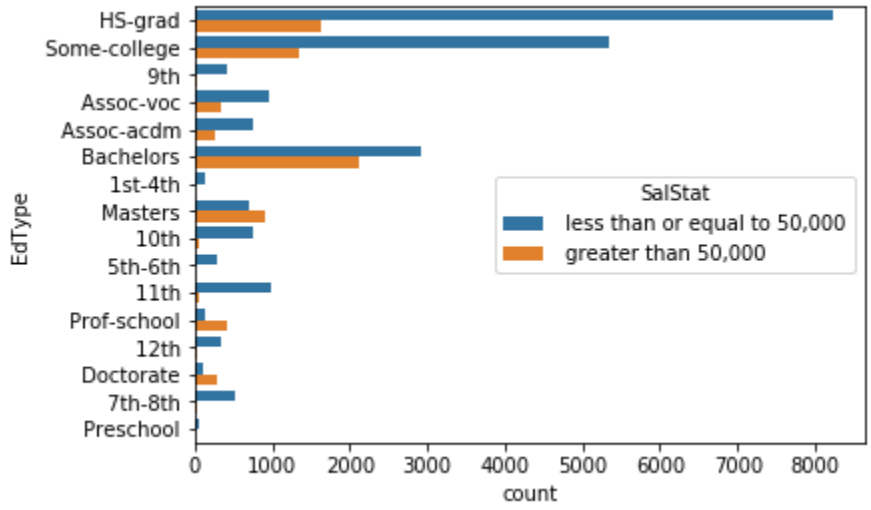


SalStat	greater than 50,000	less than or equal to 50,000
JobType		
Federal-gov	0.387063	0.612937
Local-gov	0.294630	0.705370
Private	0.218792	0.781208
Self-emp-inc	0.558659	0.441341
Self-emp-not-inc	0.285714	0.714286

SalStat	greater than 50,000	less than or equal to 50,000
JobType		
State-gov	0.268960	0.731040
Without-pay	0.000000	1.000000
All	0.248922	0.751078

According to the above diagram and table, 78.12% people working in private sectors, have income less than or equal to 50,000$. 55.86% people working as self-employed and income, earn greater than 50,000$.

EdType and SalStat:

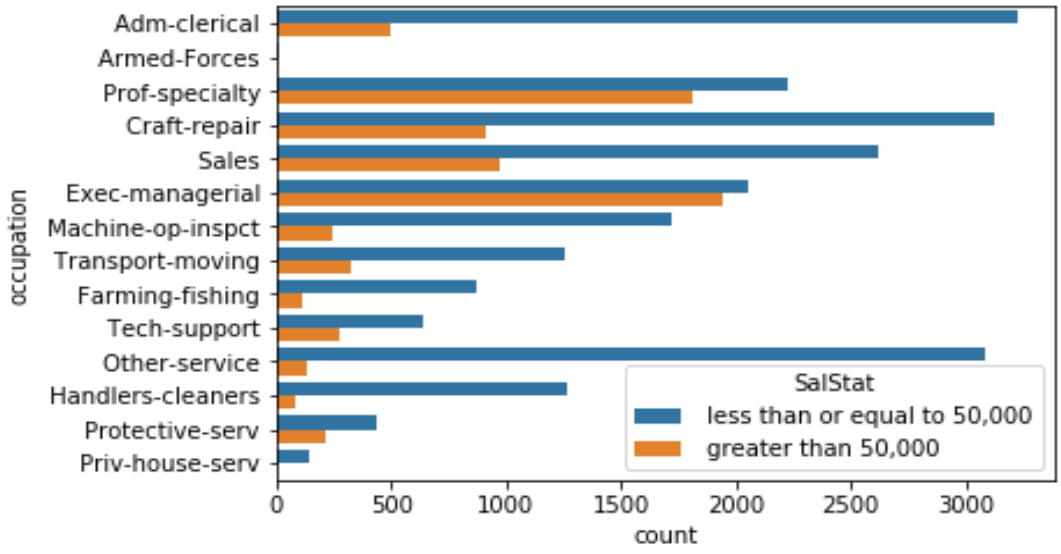


SalStat	greater than 50,000	less than or equal to 50,000
EdType		
10th	0.071951	0.928049
11th	0.056298	0.943702
12th	0.076923	0.923077
1st-4th	0.039735	0.960265
5th-6th	0.041667	0.958333
7th-8th	0.062837	0.937163
9th	0.054945	0.945055
Assoc-acdm	0.253968	0.746032
Assoc-voc	0.263198	0.736802
Bachelors	0.421491	0.578509
Doctorate	0.746667	0.253333
HS-grad	0.164329	0.835671
Masters	0.564229	0.435771
Preschool	0.000000	1.000000
Prof-school	0.749077	0.250923
Some-college	0.200060	0.799940
All	0.248922	0.751078

74.6% people in Doctorate have income greater than 50000$ per annum.
56.4% people of Masters degree have greater than 50000$ per annum.
74.9% people of school professional people have income greater than 50000$ per annum.

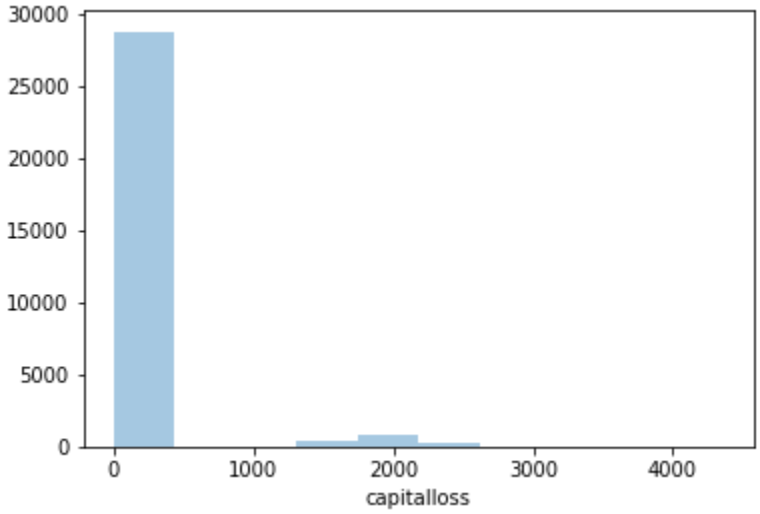
Occupation and SalStat:

SalStat	greater than 50,000	less than or equal to 50,000
occupation		
Adm-clerical	0.133835	0.866165
Armed-Forces	0.111111	0.888889
Craft-repair	0.225310	0.774690
Exec-managerial	0.485220	0.514780
Farming-fishing	0.116279	0.883721
Handlers-cleaners	0.061481	0.938519
Machine-op-inspct	0.124619	0.875381
Other-service	0.041096	0.958904
Priv-house-serv	0.006993	0.993007
Prof-specialty	0.448489	0.551511
Protective-serv	0.326087	0.673913
Sales	0.270647	0.729353
Tech-support	0.304825	0.695175
Transport-moving	0.202926	0.797074
All	0.248922	0.751078



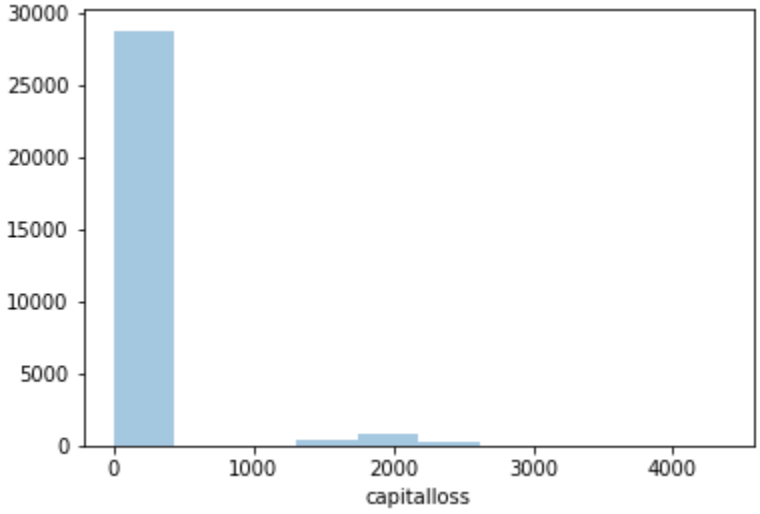
48.52% of people in Exec-managerial and 44.84% of people in prof-speciality are more likely to with income greater than 50000$ per annum. 99.3% of people in Priv-house-serv are more likely to have income less than or equal to 50,000$.

Capitalgain and SalStat:



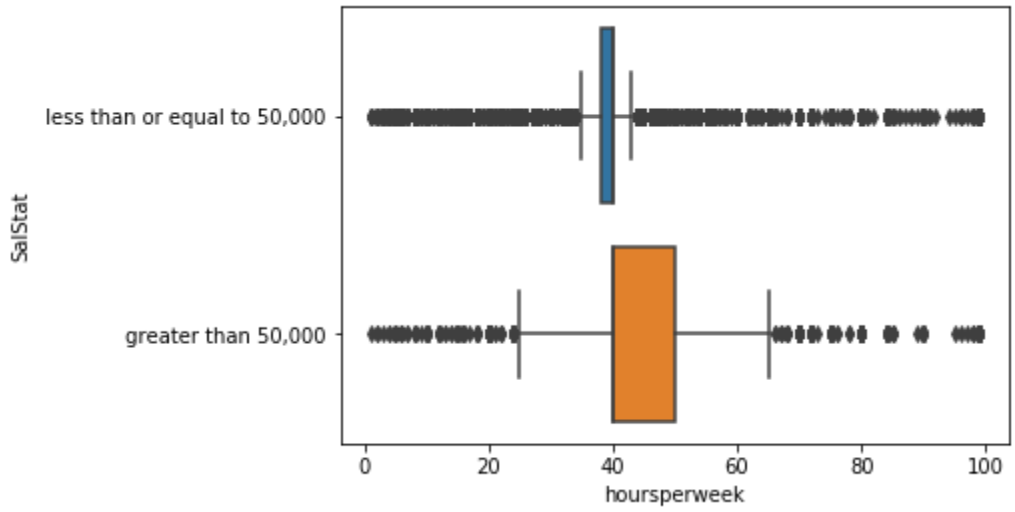
92% of people have capital gain 0 but 8% people have gained profit from selling something according to the dataset.

Capitalloss and SalStat:



95% of people have capitalloss 0 i.e. either they have not invested anything or invested but have not had any loss according to the dataset.

Hoursperweek and SalStat:



From the diagram it is clear that the people who have spent 40-50 hours per week are more likely to have salary greater than 50,000$ per annum.

Having all the experiments mentioned above, Age, JobType, EdType, occupation, capitalgain, capitalloss, hoursperweek are found as relevant columns for classification. Then dummy variables are created in place of categorical columns. Splitting independent variables and dependent variable, the whole dataset is separated in training set and testing set. Training set will help to build the model and testing set will measure the accuracy of the model.

RESULT AND DISCUSSION

After creating training and testing set, logistic regression algorithm is applied on the training set and the accuracy is checked using the testing set. The validation accuracy becomes 80.4% and confusion matrix shows the number of misclassifications 1426. The accuracy using the training set is 82.2%. Both accuracies of using training and testing set are very closed to each other.

Before removing Insignificant Variables

Model	LogisticRegression	KNeighborsClassifier
accuracy_score	0.804	0.822
Misclassified	1426	1410

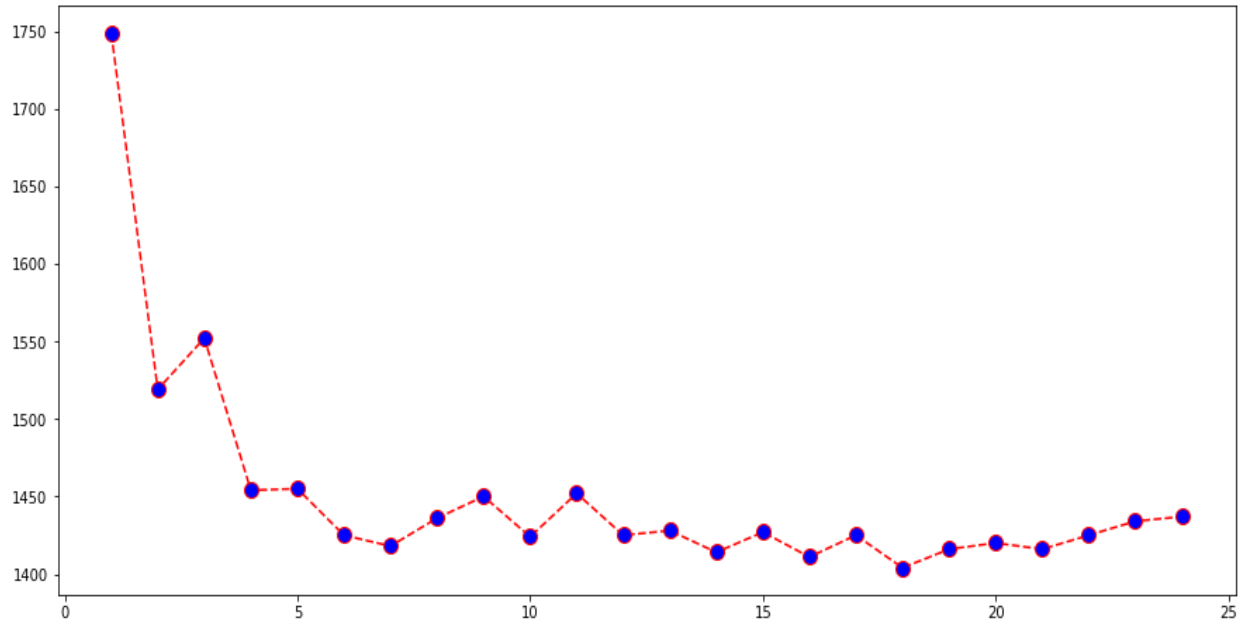
After removing Insignificant Variables

Model	LogisticRegression	KNeighborsClassifier
accuracy_score	0.838	0.843
Misclassified	1418	1400

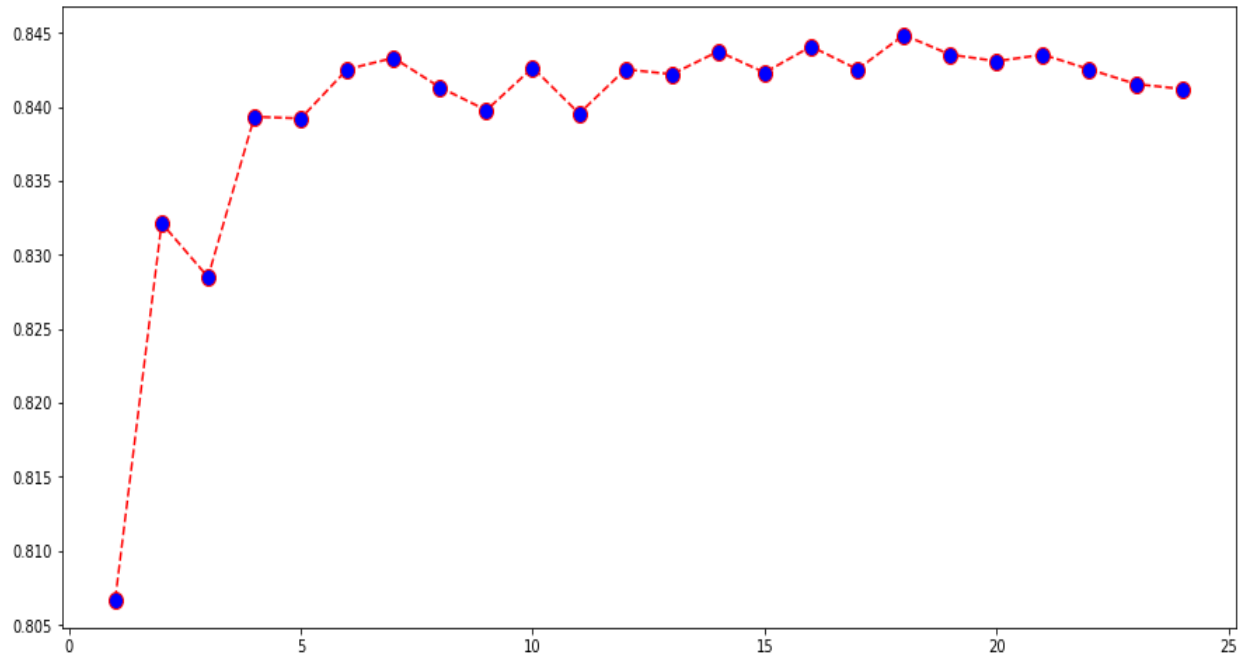
- Accuracy increases as we removed the insifnificant variables.

To improve the accuracy more, K-NN algorithm is used on the training set inputting 5 as the value of the number of neighbours in the algorithm. In this case, the accuracy using testing set becomes 83.92% and the accuracy using training set becomes 88.53%. The number of misclassifications becomes 1455. The accuracies using training set and testing set are far from each other. To improve this situation, an iteration is applied on the parameter, n-neighbors (number of neighbours) in the K-NN algorithm. This iteration shows that n-neighbors=18 (number of neighbours) is a good choice to improve the accuracy. After the iteration, the accuracy using testing set becomes 84.48% and the accuracy using training set becomes 84.3%. The number of misclassifications becomes 1404 which is better than the previous one. Both accuracies using training and testing set are closed to each other.

Plot of misclassifications: (misclassification vs n-neighbors)



Plot of validation accuracy: (testing set accuracy vs n-neighbors)



For some more improvement, Decision Tree algorithm is used on the training set inputting 10 as the value of the parameter, random state in this algorithm. The validation accuracy becomes 81.79% using testing set but the accuracy using training set becomes 97.33%. Both accuracies are very much far from each other and the accuracy using testing set is much lower than K-NN and Logistic Regression algorithms. To improve this situation, an iteration is applied on a parameter max_depth(maximum depth of the tree) where random state value is 10(fixed). The iteration shows that max_depth=9 gives the maximum validation accuracy=85.15% and corresponding training set accuracy=86.51%. In this case, the number of misclassifications is minimum and both accuracies using training and testing set are closed to each other.

Hence, Decision Tree algorithm gives the maximum validation accuracy with minimum number of misclassification because validation accuracy and number of misclassification are inversely proportional to each other.

CONCLUSION

Hence, after analysis we get the simplify data by reducing the number of variables t o be studied in the program. After applying four major classification model SVC, K-Neighbours Classifier, Decision Tree Classifier, Random Forest Classifier we come to a conclusion that Random Forest Classifier gives the best value among all of them i.e. 0.84137 and K-Neighbours Classifier, Decision Tree Classifier , SVC 0.8431957566716393 ,0.8145201392342118 , 0.7966185977125808 respectively. The random forest classifier fitted the best.

FUTURE SCOPE

For further studies, one may use some other high performance machine learning algorithms to improve accuracy level more. One may get much better result using iteration on other parameters of Decision Tree algorithm. As accuracy is inversely proportional to the number of misclassifications, higher accuracy implies better model and better model helps to take much better business decisions. But one has to be cautious about the over fitting criteria.

REFERENCES

1. Source of the data: https://archive.ics.uci.edu/ml/datasets/Census%2BIncome

2. Dataset: https://1drv.ms/x/s!AuOQjMBE9ClEdjyxMDW2kOLDi-k

3. Comparative Analysis of Classification Models on Income Prediction, by Bhavin Patel, V.Kakulapati, VVSSS Balaram, Sreenidhi Institute of Science & Technology, Yamnampet, Ghatkesar, Hyderabad, India.

4. Predicting Annual Income of Individual using Classification Techniques by Janmejay Mohanty , Stevens Institute of Technology

5. Book: Practical Machine Learning with Python, by Dipanjan Sarkar, Raghav Bali, Tusar Sharma.