Classification Algorithms for the prediction of Income from Adult Census Income Dataset

Sumit Mishra
New Delhi, India
sumit.mishra0432@gmail.com

***Abstract*—The Adult Census Income data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The data set contains the features about the people who earns more than $50K or less than $50K a year. The prediction task is to determine whether a person makes over $50K a year. Exploratory data analysis is Done on the data set to achieve insights and to get the important features then the preprocessing is done to get the data ready for the training. 70% of the data is used for training and 30% for the testing purposes. Five Classification models are trained and their performances are compared with various performance metrics like classification report and the Receiver operating characteristic curve.**

Keywords-Classification, ROC, classification report, xgb, confusion matrix, Label Encoder

I. INTRODUCTION

The Adult Census Income data was extracted from the 1994 Census bureau database. The purpose of this study is to get the insight about the income of the persons from 1994 in a particular reign and can infer about the income inequality if there's any from the data set and to successfully predict whether a person make $50K a year or not. This Adult Census data set contains the features about the people who earns more than $50K or less than $50K a year. Some features are education of the person, their nationality and age. In this project the Exploratory data analysis is done and various insights are concluded from EDA. The feature engineering is done to prepare the data for the Machine Learning models and the feature importance bar graph is plotted for the perspicacity of the features, then the Classification models are trained and the performances are compared with various performance metrics. The classification algorithms that are used are Logistic regression, Random forest Classifier, Gradient Boosting, Bernoulli Naive Bayes and the Support vector Classifier. The Best performing model is finalized for the predictions.

II. DATASET & EDA

A. Dataset

The Adult Census data set is of shape (32561,15) i.e. It has 15 features and 32,561 entries. The features are age, work class, education, education.num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country, fnlwgt and income. Income is the target Variable that is to be predicted that has two categorical values ">50K" and "<=50K". The education and education.num are related as the education.num is the numerical conversion of the education column which consist of categorical values. The relationship features tells the relationship status about the person if they are Unmarried or married in various category. The race feature is used to tell the race of the person like black, white, Asian etc. The Age column has values between 16 and 100. The fnlwgt (final weight) feature is the weight on the Current Population (CPS) files are controlled to independent estimates of the civilian non institutional population of the US. These are prepared monthly by Population Division at the Census Bureau. Three sets of controls are used for this these are:

1. A single cell estimate of the population 16+ for each state.
2. Controls for Hispanic Origin by age and sex.
3. Controls by Race, age and sex.

B. Exploratory Data Analysis

The exploratory data analysis is done to get the insight about the data and get to know about the feature dependency.

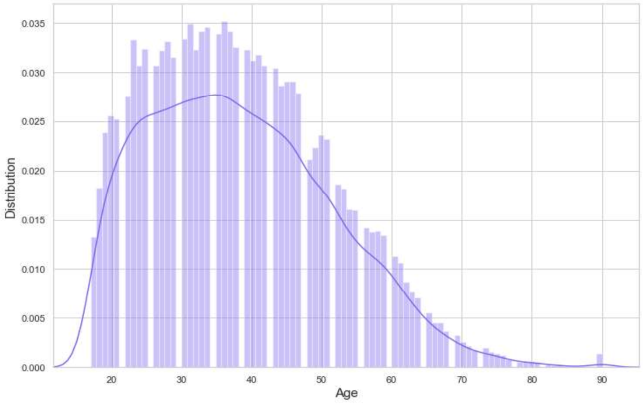


Fig – 1.0

The fig-1.0 is a distribution plot which shows the distribution of age among the whole data set. It can be inferred from the plot that the minimum age is 17 and the maximum age is 90. The distribution of age in the data set after 70 year is fairly lower.

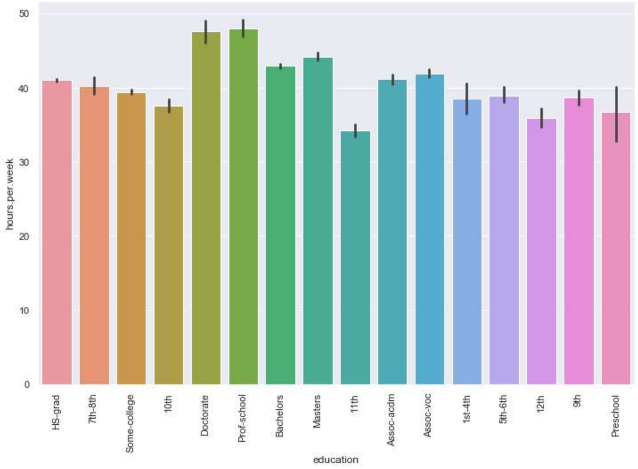


Fig – 2.0

The fig-2.0 (bar graph) shows the average hours per week according to the education of the person. It can be seen that Prof-School and Doctorate have the highest hours per week but as it's the average, it can be seen that the person with the preschool education also have an average hours per week of 36.64.

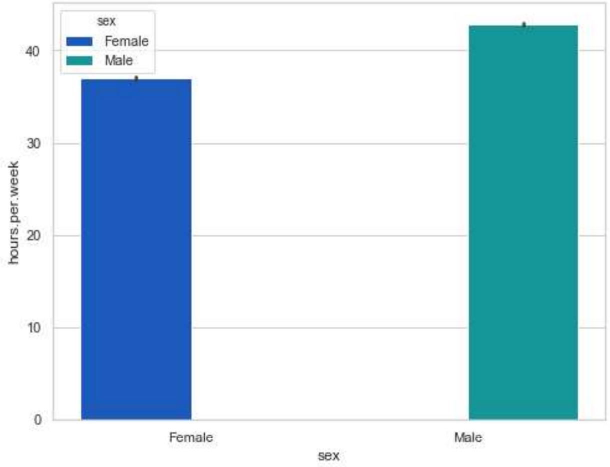


Fig – 3.0

The fig-3.0 (bar graph) shows the average hours per week between female and male .As it can be interpreted from above, for females average hours per week is 36.4 hours whereas for male it is 42.4 hours.

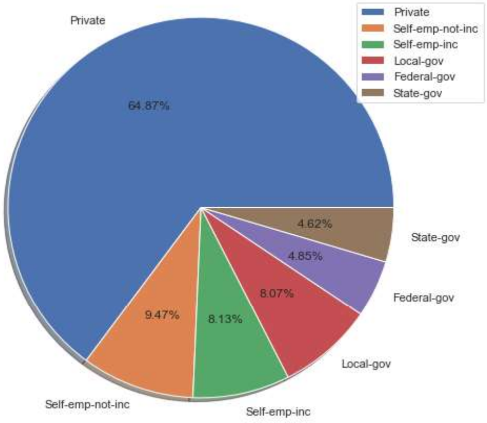


Fig – 4.0

The fig-4.0 (pie chart) shows the people who earns more than $50K segregated by their work class which are Private, self-employed not Inc., self-employed Inc., local government, federal government and state government. The majority of the people who earns more than $50K work in the private class, about 64.87% and then comes the self-employed not Inc. with 9.47%. It can be assuredly interpreted that theirs a disparity between private and other work classes. The work class with the minimum percentage of people who earns more than $50K is State Government with 4.62%. This pie chart clearly shows that the Private Companies pays more than the other ones by a huge proportion.

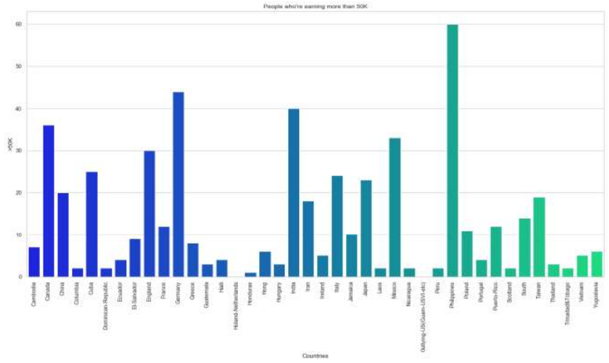


Fig – 5.1

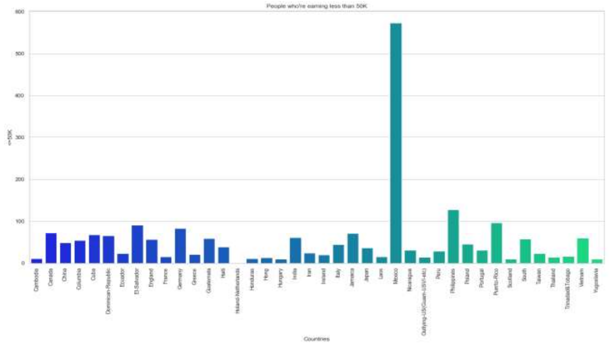


Fig – 5.2

The fig-5.1 and fig-5.2 bar graphs shows the people who earns more or less than $50K grouped according to their nationality. This bar graph is plotted excluding the united nations as there's a disparity between United nation and other nation. In the first bar graph we can clearly see that people who are earning more than $50K after US are from Philippines and Germany. and the for people earning less than $50K after US are from Mexico.

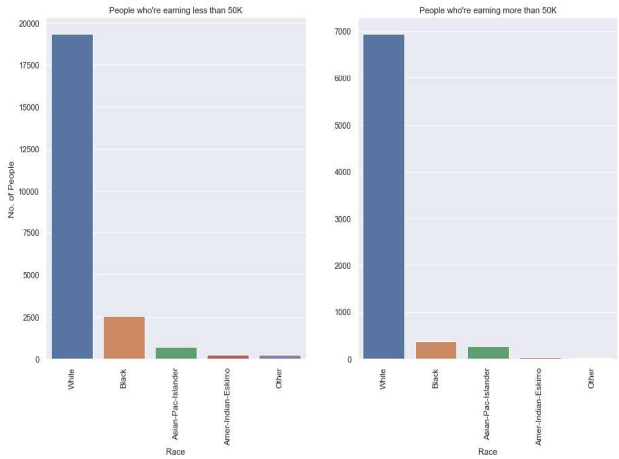


Fig – 6.0

The fig-6.0 shows the disparity between the incomes of the peoples according to their races. It can be inferred from the above graph that most people who are earns less or more than $50K are white people and followed by the Black People and at last comes the Other category.

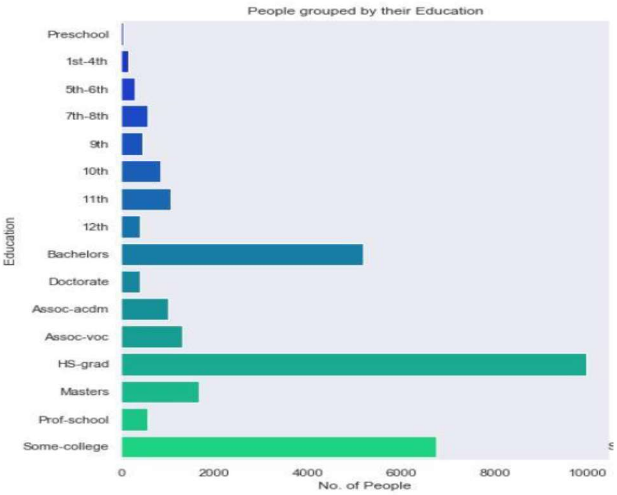


Fig – 7.1

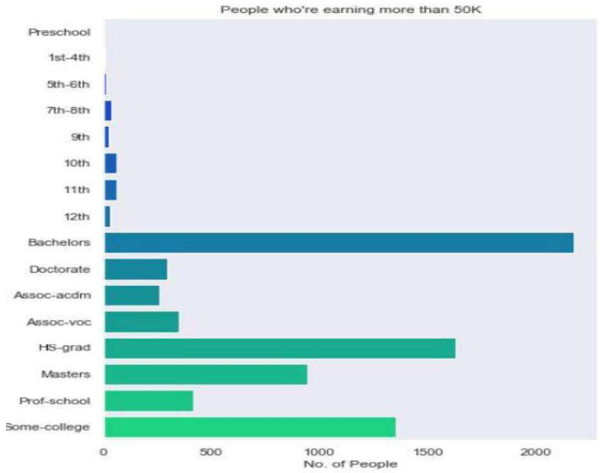


Fig – 7.2

The fig-7.1 and fig-7.2 are two horizontal bar graphs in which the first one shows the distribution of people according to their Education and the second one shows the people earning more than $50K a year according to their education. It can be inferred that the most people are from HS-grad i.e. 10,501 and then some colleges with 7,291 and least people are from Prep School. But from the second bar graph (fig-7.2) it can be clearly seen that most people i.e. 2,221 who earns more than $50K are Bachelors then the HS-grad with 1,675 people and 1,387 people from some colleges. No person from prep school earns more than $50K.

III. PREPROCESSING AND FEATURE ENGINEERING

Adult Census Dataset contains 15 Features.The features are age, work class, education, education.num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country, fnlwgt and income. Income is the target variable which is to be predicted.

All the duplicate rows are deleted from the dataset. The dataset also contains some missing values ("?") in the columns that are work class, occupation and native country. And the amount of missing values is just 5.6% of the dataset hence they are dropped.

A looped Label Encoder is used to convert all the categorical values in every column to the numerical values for the training and testing. The features that has categorical values are education, marital status, occupation, relationship, race, sex, native country and income. The target variables are 1 and 0 only. The education.num column is dropped as it's same as the education column which is converted to the numerical variable. The processed data frame is then shuffled before splitting it into training and test data.

A. Correlation matrix heatmap

A heatmap of the correlation matrix has been plotted to check the correlation between the features, if there's a positive or negative correlation.

The fig-8.0 is the heatmap of the correlation matrix of the features. It can be inferred that the education.num has a fair positive correlation with hours per week and the capital gain. The capital loss and the capital gain has a negative correlation.

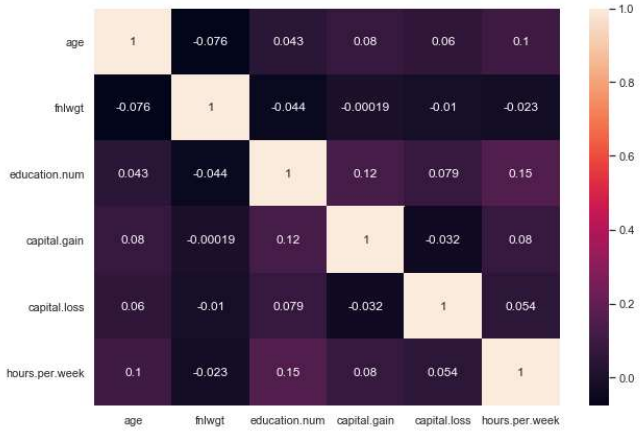


Fig – 8.0

B. Feature Importance

Gradient Boosting is used to plot the feature importance bar graph. From this feature importance graph as shown in the fig-9.0, it can be inferred that the Capital.gain is the most important feature which is followed by the Age. Capital.loss and education have the same importance for this binary classification. Native country has the least importance as compared to the other features.

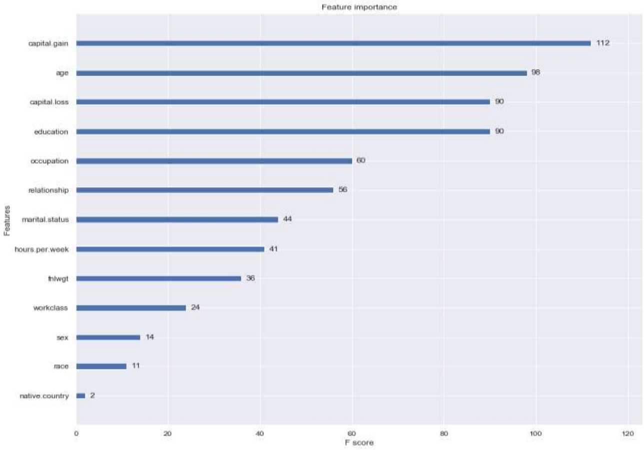


Fig – 9.0

IV. Splitting Dataset and modelling

The shape of the dataset after the deletion of the duplicates is (32537, 15). The dataset is split where 70% is the used for training the model and 30% for testing the model. Hence out of 32,537 data entries, 21,485 are used for training and 9,209 are used for testing the model. 5 Classification Algorithms are used. The algorithms and their hyper parameters are given below.

A. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression). Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

LogisticRegression (C = 0.5, max_iter = 500)

B. Random forest Classifier

Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. It's a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

RandomForestClassifier (n_estimators = 200)

C. Gradient boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. gradient boosting combines weak "learners" into a single strong learner in an iterative fashion.

XGBClassifier (learning_rate = 0.35, n_estimators = 200)

D. BernoulliNB Classifier

BernoulliNB implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, Boolean) variable. Therefore, this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a BernoulliNB instance may binarize its input (depending on the binarize parameter).

BernoulliNB (alpha = 0.3)

E. Support Vector Classifier

A Support Vector Classifier (SVC) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data (*supervised learning*), the algorithm outputs an optimal hyper plane which categorizes new examples. In two-dimensional space this hyper plane is a line dividing a plane in two parts where in each class lay in either side.

SVC (kernel = 'rbf', max_iter = 1000, probability = True)

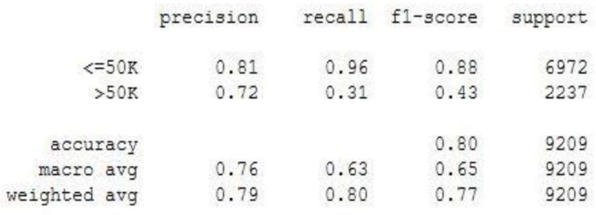
V. Model Comparison

The Classification Report and the accuracy of Each Machine Learning model is given and the ROC Curve is also plotted to compare the performance of these classification Algorithms.

A. Classification Reports

The Classification Reports of the machine learning models are given:

Logistic Regression:



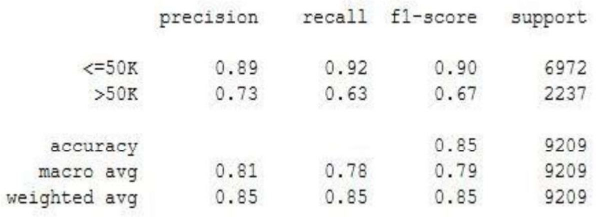
```
              precision    recall  f1-score   support

       <=50K       0.81      0.96      0.88      6972
        >50K       0.72      0.31      0.43      2237

    accuracy                           0.80      9209
   macro avg       0.76      0.63      0.65      9209
weighted avg       0.79      0.80      0.77      9209
```

Logistic Regression is the first model trained and it has an accuracy of 80.18%.

Random Forest Classifier:



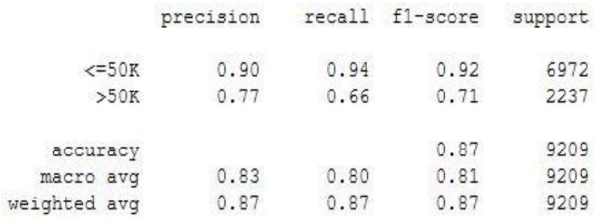
```
              precision    recall  f1-score   support

       <=50K       0.89      0.92      0.90      6972
        >50K       0.73      0.63      0.67      2237

    accuracy                           0.85      9209
   macro avg       0.81      0.78      0.79      9209
weighted avg       0.85      0.85      0.85      9209
```

Random Forest Classifier also has a very good accuracy of 85.25%. It also has a very good precision score for both the classifications.

Boosting Gradient Descent:



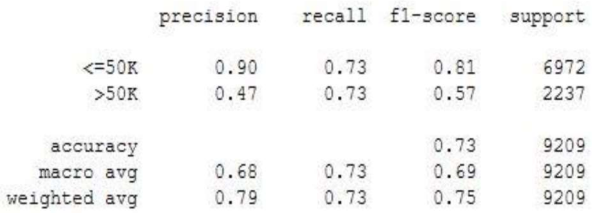
```
              precision    recall  f1-score   support

       <=50K       0.90      0.94      0.92      6972
        >50K       0.77      0.66      0.71      2237

    accuracy                           0.87      9209
   macro avg       0.83      0.80      0.81      9209
weighted avg       0.87      0.87      0.87      9209
```

Boosting Gradient Descent have the highest accuracy of them all of 86.99%.

BernoulliNB:



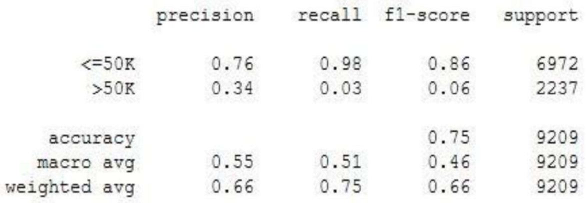
```
              precision    recall  f1-score   support

        <=50K       0.90      0.73      0.81      6972
         >50K       0.47      0.73      0.57      2237

     accuracy                           0.73      9209
    macro avg       0.68      0.73      0.69      9209
 weighted avg       0.79      0.73      0.75      9209
```

BernoulliNB has an accuracy of 73.25%.

Support Vector Classifier:



```
              precision    recall  f1-score   support

        <=50K       0.76      0.98      0.86      6972
         >50K       0.34      0.03      0.06      2237

     accuracy                           0.75      9209
    macro avg       0.55      0.51      0.46      9209
 weighted avg       0.66      0.75      0.66      9209
```

The last model trained was the Support Vector Classifier has an accuracy of 74.98%.

B. ROC Curve

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or *probability of detection* in machine learning. The false-positive rate is also known as the fall-out or *probability of false alarm* and can be calculated as (1 − specificity).

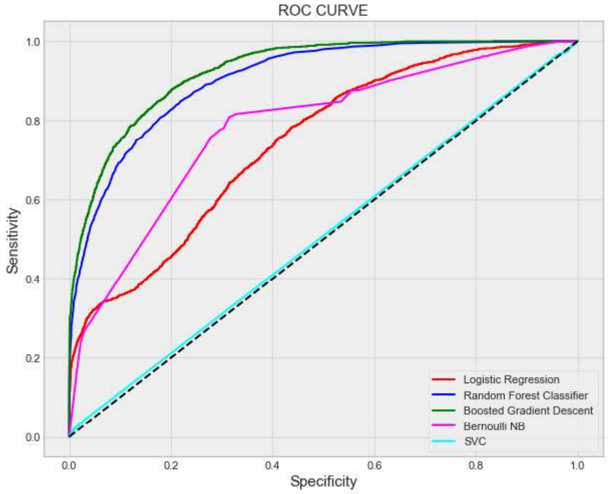


Fig – 10.0

From this ROC Curve in Fig-10.0 it can be inferred that the Boosting Gradient is performing the best as the area under the curve is the highest. The Random Forest is also a good performing Algorithm after the boosting gradient. The SVC is the worst performing model as it just shadowed the random guess line which is the worst-case scenario.

VI. Conclusion

The main aim was to successfully predict the income of the person from dataset using various classification algorithms and compare their performances. The Exploratory Data Analysis is done on the data set to get insight about the data and get the correlations of the features present. The importance of the features is also plotted. The preprocessing and feature engineering is done and five Classification Algorithms are trained. The performances of the models are compared on the basis of classification report and ROC curve. The Boosting Gradient Descent comes to be the best performing algorithm above all other models with an accuracy of 86.99% and over all generalizing well.

A. Authors and Affiliations

Sumit Mishra

References

[1] Kaggle Adult Census Income Data Set - https://www.kaggle.com/uciml/adult-census-income

[2] Logistic Regression - https://www.statisticssolutions.com/what-is-logistic-regression/

[3] Random Forest Classifier - https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[4] Boosting Gradient Descent - https://en.wikipedia.org/wiki/Gradient_boosting

[5] Bernoulli NB - https://scikit-learn.org/stable/modules/naive_bayes.html

[6] Research Paper Template- https://www.ieee.org › web › org › conferences › Conference-template-A4

[7] Receiver operating Characteristic Curve - https://en.wikipedia.org/wiki/Receiver_operating_characteristic