Income Predicion using Machine learning

Nikhil Dharmavaram, MSc Big Data Analytics and Artificial Intelligence, Letterkenny Institute of Technology, l00157103@student.lyit.ie
Under the supervision of Dr.Shagufta Henna, Letterkenny Institute of Technology, Letterkenny, Co. Donegal

***Abstract*— In many countries leading imbalance of income and wealth is the biggest concern especially in developed countries like the United States. One compelling reason to reduce the world's rising level of economic disparity is the possibility of reducing poverty. The theory of fundamental moral equality promotes long-term growth and improves the country's economic stability. There are many attempts by countries and their governments trying to address this issue to provide an optimal solution. This paper aims to show how machine learning algorithms like logistic regression, decision tree and random forest can be applied to predict whether a person makes more than 50K dollers or not. The dataset used for this study is Adult UCI dataset. This big data project starts with data selection, data loading, data cleaning and preprocessing, exploratory data analysis, visualization and model building using machine learning. Random forest has the best areaUnderROC value and thus the model prediction is based on random forest classifier..**

***Index Terms*— Machine learning, logistic regression, decision tree, random forest, areaUnderROC, exploratory data analysis, visualization**

I. Introduction

HUmans have been increasingly reliant on data and information in society over the past two decades, and as a result, systems for data collection, interpretation, and retrieval on a large scale have advanced. Data Mining and Machine Learning have not only used them for intelligence and exploration, but also to uncover latent trends and hypotheses that have contributed to the prediction of difficult-to-predict future events. In recent years, the issue of economic growth has been a major source of concern. Making the poor's lives happier is not the sole criterion for combating this issue. This big data analytics project can be used to study various patters that effect the income of a person like age, occupation and education[1]. The adult uci dataset was used for this project which was fairly clean so there was not much cleaning needed and the dataset was ready for next process with a bit of cleaning. There are many insights which can be unearthed by doing data analytics and visualization of which are discussed in the next sections and to end it all the project is finished by building a model which predicts the income of a person is more than 50k dollars using three binary classification algorithms while decision tree and random forest can also be classified under multiclass classification. The data is first split into training and test data in 70:30 ratio and then applied to these algorithms to get the ideal output[2]. Finally the result is displayed with income prediction with respect to age and income prediction with respect to occupation.

II. Dataset Overview

The data set we are using is called adult UCI data. This data set contains the mixture of both categorical and numerical values. The numerical value columns are age, final weight, capital gain, capital loss and hours per week and categorical value column include work class, education, marital status, occupation, relationship, race, gender and native country.

Categorical Attributes

1 work class- this column contains the work category of individuals like private self-employed, federal government, without pay, never worked etc.

2 education- this column contains the highest education of the individual for example preschool, professional school, bachelors, doctorate etc.

3 marital status- this column contains the marital status of the individual like married, divorced, separated, never married etc.

4 occupation- this column contains individuals' occupation like tech support, craft repair, sales, professional specialty, farming-fishing, transport moving, armed forces etc.

5 relationship- this column contains the individual's relation in a family for example husband, wife, on child, other relative etc.

6 race- this column shows us the race of the individual like white, black, Asian etc.

7 gender - this column contains the gender of the person.

8 native country-this column contains the individual's native country like United States, England, India, Japan etc.

Numerical attributes

1-Age-this column contains the age of the person

2- Final weight- this column contains the weights on the cps files.

3- Capital gain: contains capital gain of an individual person.

4- Capital loss-this contains the capital loss of an individual.

5- Hours-per-week-this contains the number of hours a person works in a week.

There are a total of 48,832 rows and 15 columns. There are some missing values in the data set and these missing

values are filled with '?'. The income column has only two values which is also known as class values. Values of this income column are ¡=50K and ¿50K which means it's a binary classification problem.

III. Data preprocessing and transformation:

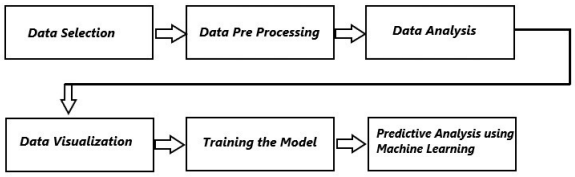

Fig. 1. Pipeline

Before we begin data preprocessing, we must check the data for any of the NaN/Null values. The adult UCI data set does not contain any null values nor missing values but there is a '?' in the columns wherever the data is missing. The '?' values are primarily in work class column, occupation column and native country column. There are two ways we can deal with this '?'.

1. Drop all the rows which have '?' In them.
2. Replace the '?' with the mode value of that column (value with the most occurrence).

As the occurrence of these '?' in columns are significantly more we implement the second option which is replacing the '?' With the mode value of that column.

• In workclass column there are a total of 2799 rows where the value is '?' and the mode value of workclass column is 'Private' with 33906 rows so the missing value will be replaced by 'Private' .

• In occupation column there are a total of 2809 rows where the value is '?' and the mode value of occupation column is 'Prof-specialty' with 6172 rows, so the missing value will be replaced by 'Prof-specialty".

• In native-country column there are a total of 857 rows where the value is '?' and the mode value of native-country column is 'United States' with 43832 rows, so the missing value will be replaced by 'United States'.

IV. Feature Engineering:

Feature engineering is a process of applying operations on rows or columns with already existing values to make it easier to interpret or make it easier to feed it to machine learning.

We use 'count' function to get all the distinct values of a particular column. Marital status column and education column in particular have many number of values that are very hard to interpret and can be segregated under a more general value.

Marital status contains seven distinct values namely Married-civ-spouse, Never-married, Divorced, Separated, Widowed, Married-spouse-ab and Married-AF-spouse.

Marital-Status	count
Married-civ-spouse	22379
Never-married	16117
Divorced	6633
Separated	1530
Widowed	1518
Married-spouse-ab...	628
Married-AF-spouse	37

Fig. 2. Marital Status before Feature Engineering

This can be lowered for easy usage and easy interpretation by using regexp-replace function we can replace all the redundant values with more general value.

Just like marital status education column contains huge number of distinct values in it like HS-grad, Some-college, Bachelors, Masters, Assoc-voc, 11th, Assoc-acdm, 10th, 7th-8th, Prof-school, 9th, 12th, Doctorate, 5th-6th, 1st-4th and Preschool which totals to 16 different values. This can be reduced to avoid confusions and for easy interpretation as High-School, Some-Degree, Bachelors, School, Masters and Doctorate.

Marital-Status	count
Married	22416
Un-Married	16117
Others	10389

Fig. 3. Marital Status after Feature Engineering

V. Data Exploration/Visualization:

Data visualization is the graphical representation of a given dataset. Utilizing graphs, charts and maps data visualization gives good insights, trends and patterns in the dataset. As this is the age of big data, data visualization has become a most important tool to represent millions of rows of data by making it easier to understand and highlighting only important trends and patterns.

From our dataset we can answer six most important question which affects the income of a person. Then we will try to get the answers from visualizing the dataset into bar graphs and line graphs.

1. Does education play a major role in salary and what is the minimum level of education needed to ensure a high salary?

As we can clearly see from fig-3 that as the education level of the person increases the chances of getting higher salary increases.The minimum level of education required to ensure income more than 50K is 'Some-Degree'.

2. Will marital status affect the salary of a person?

We can clearly observe from fig-4 that marital status of a person heavily affect his salary as the chances of him getting more salary has tripled compared to unmarried person.

3. With all other factors being the same will sex of a person determine him or her getting a higher salary?

Yes as observed in fig-5 the gender do affect the person in

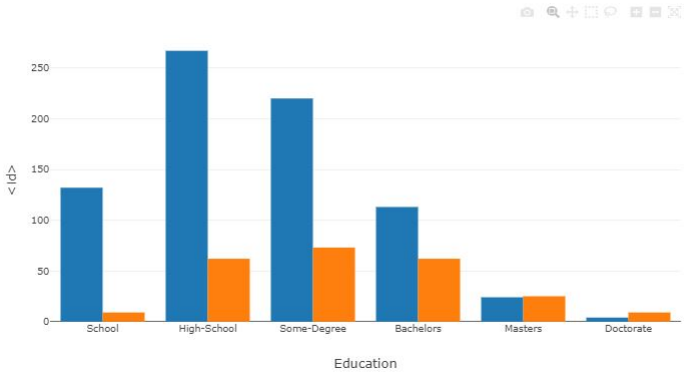

Fig. 4. Education

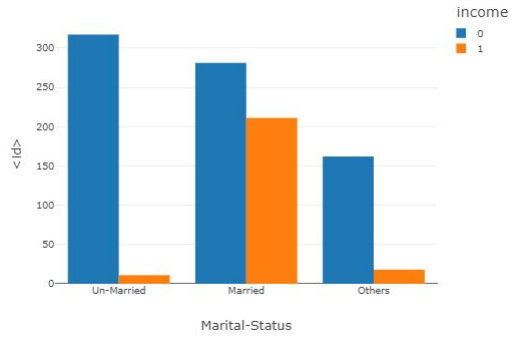

Fig. 5. Marital-Status

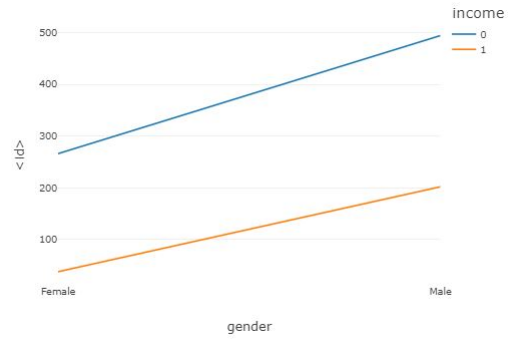

Fig. 6. Gender

getting higher salary.

4. Will the age of a person play a significant role in defining his salary?

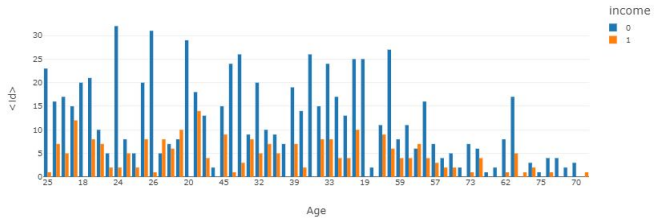

Fig. 7. Age

From fig-6 we can observe that a person tends to get higher salary when he/she is in his/her 20's and gradually decreases as the person ages and the chances of earning more salary almost perishes when the person crosses 65 years.

5. Will the race of a person be a significant factor in defining his salary?

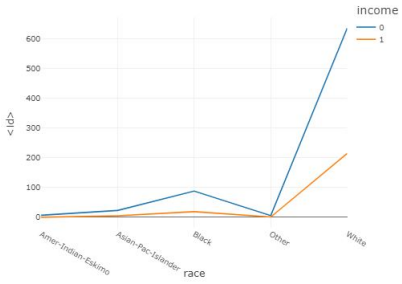

Fig. 8. Race

From fig-7 we can notice that in this dataset the majority of the people belong to white background and thus the people with white race has the chances of higher salary.

VI. Big Data Analytic Architecture And Tools

For the implementation of this project spark dataframe was used and databriks was used as the platform.

The main libraries it boasts are the following: Spark SQL,

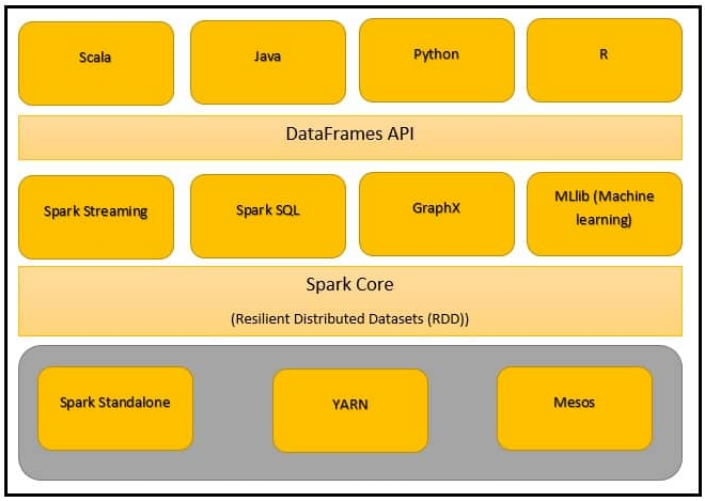

Fig. 9. Spark Architecture

MLlib, GraphX and SparkStreaming. Even though the implementation was in spark the data was converted to pandas to find out the correlation of the different numeric attributes and used it for visuaization.

A. Code List

The project was implemented using spark data frame in technical-project-nikhild-l00157103.ipynb which contains all the processes from dataset overview, data loading, data cleaning ,data pre-processing, feature engineering, data visualization and building machine learning algorithms.

VII. Machine Learning Models:

Before applying machine learning algorithms data pre processing is required so as to convert all the categorical values to numerical values.

1. Category indexing:

This process is assigning a numerical values to all the category from 0, 1, 2 etc. This method is suitable for ordeal variables.

2. One-Hot Encoding:

One hot encoding converts the categorical values into binary vectors with a max of 1 non zero value. In this dataset we have both ordinal and nominal variables like education column and relationship column respectively. One Hot Encoding is used to convert all the categorical values into binary vectors. There are high chances of increase in accuracy by converting the categorical values using correct method.

In this project StringIndexer and OneHotEncoder was used to convert categorical values to numeric. Since there is more than one stage of transformation we can tie all the levels of transformation in a pipeline which results in simplification of our code.

After converting the values the dataset split into training and test data in 70:30 ratio. After splitting the initial model was created using param grid and five fold cross validation using the training set and finally we use BinaryClassificatioEvaluator to evaluate our models.

The following machine learning algorithms were constructed and demonstrated

A. Logistic Regression

Logistic regression is mainly preferred to use as regression analysis when the dependent variable is binary. Logistic regression is a predictive analysis. It is a technique that is used to show relationship between one dependent binary variable and one or more independent variable by estimating probabilities using a logistic function. The logistic function is a S shaped curve that takes real valued variable and maps it between 0 and 1.logistic regression similar to linear regression uses an equation as representation. First an input value x is selected and it is mapped linearly using weights to predict the output y. The difference between linear regression and logistic regression is that logistic regression outputs 0 or 1 unlike numerical values in linear regression. Logistic regression models the probability of the default class. For example from the dataset we are modeling people's income as more than 50k or less than equal to 50k given their age, occupation etc, then the first class id income and the logistic regression model is given as probability of income more than 50k given age/occupation. Using maximum likelihood estimation coefficient value from our training data is obtained to pass it through logistic regression[3]. In the project LogisticRegression library was imported from pyspark.ml.classification and then an object was created using the following parameters (labelCol="label", featuresCol="features", maxIter=10).The linear regression model was supplied with training data to carry out the operations.

B. Decision Tree

A decision tree is a graphical representation of a decision based on some condition which maps all the possible solutions. It is called as decision tree because it starts with one node(root) and then branches on both the sides to map all the possible solutions and last node is called as leaf node. Decision tree is a supervised learning algorithms but it can be used to solve regression and classification problems also. The decision tree learns from simple decision rules and applies it on training data to predict the value of a training variable. In decision tree to predict the income of our dataset we need to start from the first node i.e. root node and then compare the values with the attributes in the dataset. The process includes comparison and then jumping to next node if the values don't corresponds. Based on our dataset the decision tree uses CART-classification and regression tree and this is selected based on type of target variable[4]. In this project DecisionTreeClassifier library was imported from pyspark.ml.classification and then a tree model was created giving number of nodes and depth as parameters. Decision tree is plotted and then BinaryClassificationEvaluator is used to evaluate the decision tree

C. Random Forest Classifier

Random forest is also known as Ensemble Learning as it integrates the output of various models rather than one to get a more accurate output. Because of its simplicity and high accuracy Random forest is one of the most popular algorithms for regression. Random forest is a supervised machine learning algorithm. Random forest generates n decision tree estimators which in scikit learn defaults to 100 where it is called n-estimators. Each tree is given a sample of the full data set with replacement. CART trees are created using the data points and features which was used to create the tree to the maximum depth possible using the feature of a subset[5]. The distance matrix is created from the data points using dissimilarity measure. To create the final cluster portioning around method clustering operaton is carried on this dataset taking number of clusters as inputs.

In this project RandomForestClassifier library was imported from pyspark.ml.classification and BinaryClassificationEvaluator is used to measure the accuracy of the model before the training data is fed into 5 fold cross validation process.

VIII. Results and analysis

The ROC value was calculated before feeding the data to the algorithms.

The trining and test data was fed to 3 supervised machine learning algorithms

1. Logistic Regression
2. Decision Tree
3. Random forest

Random forest gave us the best accuracy compared to other two algorithms.

The accuracy of logistic regression was 88 percent and the accuracy of decision tree was 76 percent and the accuracy for random forest was 89.12 percent.

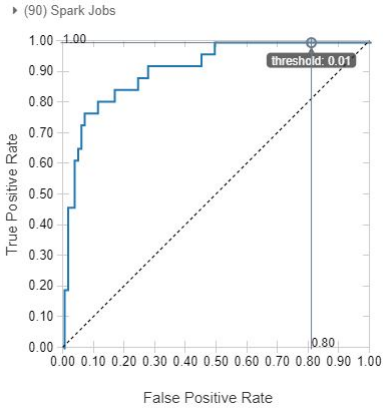

Fig. 10. ROC

Random forest was used to make the predict if a person gets more than 50K or less than equal to 50K because the areaUnderROC value of random forest is high compared to other two algorithms.

1.Prediction of Income based on Age

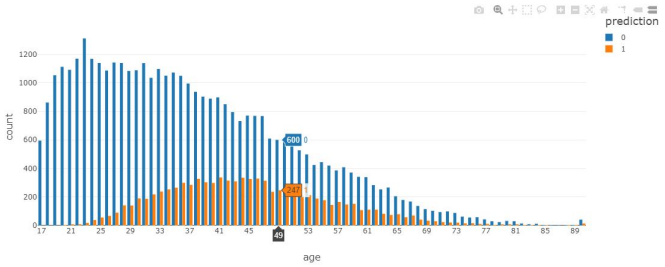

Fig. 11. Prediction based on Age

From prediction we can clearly see that age is the biggest factor in determining the income of a person. A person in his 20's and 30's earns more and as the age increases the income decreases.
A person earns more than 50K a year when the person is between when he is 30 years till he is 50 years and then slowly the downward shift of the graph can be observed.

1.Prediction of Income based on Occupation

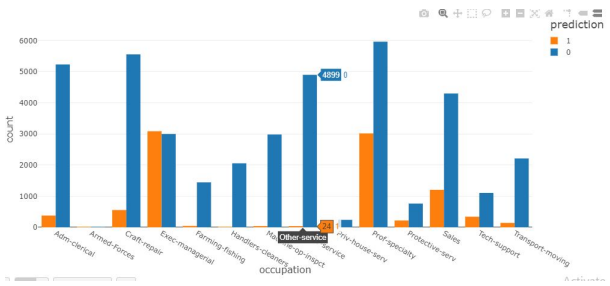

Fig. 12. Prediction based on Occupation

From prediction we can observe that the occupation of a person is the most important factor while determining the income of a person. Exec-Managerial job has the highest chances of getting income more than 50k as the people who get less than 50K are lower than those people who get more than 50K
Other services, farming-fishing and handlers-cleaners have the lowest chances of getting income more than 50k

IX. Conclusion

Adult UCI dataset was selected to apply for big data analytics project. The dataset was quite clean so much cleaning and preprocessing was not required. The cleaned dataset visualized to get some more insights using bar graphs and line graph. Then the data was split 70: 30 into training data and test data to be applied for three machine learning models. Logistic regression accuracy was quite good with 88 percent along with 79 percent of decision tree and 89 percent of random forest.as the Accuracy and areaUnserROC scores of random forest was good the data was tested with random forest model and the results were surprisingly good. Key findings are People whose age is between 25 and 55 have higher chances of earning well compare to others and people who work in Exec-Managerial and Prof-Speacility have the chances of earning more than 50K compared to other professions.

X. References

[1] Vidya Chockalingam, Sejal Shah and Ronit Shaw: "Income Classification using Adult Census Data", https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a120.pdf.

[2] Sisay Menji Bekena:"Using decision tree classifier to predict income levels", Munich Personal RePEc Archive 30th July, 2017

[3] Mohammed Topiwalla: "Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting", University of SP Jain School of Global Management.

[4] Alina Lazar: "Income Prediction via Support Vector Machine", Inter-national Conference on Machine Learning and Applications - ICMLA 2004, 16-18 December 2004, Louisville, KY, USA.

[5] S.Deepajothi and Dr. S.Selvarajan: "A Comparative Study of Classifica-tion Techniques On Adult Data Set", International Journal of Engineering Research Technology (IJERT), ISSN: 2278-0181 Vol. 1 Issue 8, October-2012.