INCOME PREDICTION USING MACHINE LEARNING

Mr.Venkata Abhilaash Annamreddi[1] **G01336658** *Student, Data Analytics Engineering, George Mason University ,Fairfax,VA,22030, Graduate Student, vannamre@gmu.edu*[1].

Abstract: - Machine learning is a technology which allows a software program to become more accurate at pretending more accurate results without being explicitly programmed and also ML algorithms use historic data to predict the new outputs. This paper gives you an idea about how the data is useful for identifying the income of a particular group based on various input variables like whether they were married, Hrs. of work etc. and the main goal of this paper is to explain the importance of each machine learning algorithms and when to use the particular algorithm and gives you an idea on supervised and unsupervised learning and some explanation on classification algorithms. Basic guidelines for the preparation of a technical work for predicting an accurate machine learning model. Classification is one part of supervised learning. By the way supervised learning is predicting the model on particular expected output with both input and output variables. In contrast to the supervised there is another one called unsupervised learning which focuses on getting some pattern without any predicting variables by your side all you have is only data.

Keywords: - *Machine Learning, Logistic regression, Decision tree,Random forest,Model selection, Supervised Learning*

I INTRODUCTION

In the normal or in the upcoming future,one of the reasons for an employee switching companies is the salary of the employee.

Employees keep on switching companies to get the expected salary. And this leads to loss of the company and to overcome this loss I come with an idea what if I develop a phenomenon where an employee gets the desired salary from the Company or Organization. In the current Competitive world everyone has high expectations and goals. But we cannot randomly provide everyone their expected salary. There should be a system which should measure the ability of the Employee for the Expected salary.

We may not be able to decide the exact salary ,but what we can do is predict it by using certain data sets.

In precision, a prediction is an assumption about a future event.

In this paper the main aim is predicting the salary column and making a suitable model to predict it and how the salary is influenced by the different variables like marriage,education,type_employer etc . So that an Employee can get the desired salary on the basis of his qualification and his status in current factors that are accompanying it. For developing this system,Since the predicting variable is Categorical we are gonna use Logistic Regression,Decision tree,RandomForest and will finalize the model based on accuracy and predicted output,

The machine learning algorithm we are gonna apply falls under Supervised learning. It is basically a learning task of a learning function that maps an input to an output of a given example. In supervised learning each example is a pair having input parameters and the desired output value.

Logistic regression algorithm in machine learning is a supervised learning technique to approximate the mapping function to get the best predictions. The main goal of regression is the construction of an efficient model to predict the dependent attribute from a bunch of attribute variables. A regression problem is when the output value is of categorical type like salary in our case..

II LITERATURE REVIEW

1) Susmita Ray," A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT Con), India, 14th -16th Feb 2019 a brief review of various machine learning algorithms which are most frequently used to solve classification, regression and clustering problems. The advantages, disadvantages of these algorithms have been discussed along with comparison of different algorithms (wherever possible) in terms of performance, learning rate etc. Along with that, examples of practical applications of these algorithms have been discussed.[1]

2) Salary Prediction Using Machine Learning by Prof. D. M. Lothe1, Prakash Tiwari, Nikhil Patil, Sanjana Patil , Vishwajeet Patil, design of a novel prediction engine for predicting suitable salary for a job" | volume 6 || issue 5 || may 2021 || issn (online) 2456-0774 international journal of advance scientific research and engineering trends is focussed on the major requirement for employee is salary the sole purpose of this paper is to help them identify which salary is apt and sufficient to particular employee so that he won't indulge any losses to the company or Organization.[2]

3)Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study Ignacio Mart´ın * , Andrea Mariello , Roberto Battiti , Jose Alberto Hernandez ´ Accurate recruitment of employees is a key element in the business strategy of every company due to its impact on companies' productivity and competitiveness.As a result, this dataset, which covers a period of 5 months, includes only ≈ 4,000 job posts, which are represented as vectors of ≈ 2,000 features.The dataset under study comprises 3,970 job posts. In order to gain useful insights into the online job recruitment for IT professionals, they compare and contrast different strategies and machine learning models which will be discussed throughout the paper..[3]

4) Job Salary Prediction Archit Khosla This dataset is part of a challenge in kaggle The average salary in the UK is 31,554.441 units. The maximum salary they observed is in 'Worle' and 'Denver' with the value being 153,600 and 100,000 respectively. However, the minimum salary being in the 'Northfield' and 'Llanrug' which is just 5,088 units. One surprising finding was that the average salary in 'Channel Islands' is about 71,500 which is about 40,000 above average; however, the rest of the places are about 10,000 above or below average.They developed a small table at the end with the RMSE values for each model and they found the highest is from model1 i.e absolute error mean absolute error mean square error model 0 796966682.3 4341.344633 321230055.7 Model 1 737971741.3 4019.979416 261936172.6[4]

----------***----------

III METHODOLOGY

In order to gain useful insights into the factors that influence the salary whether it is greater than 50k or less than 50k. We compare different strategies and machine learning models. The methodology has different phases like: Data collection, Data cleaning, Data Visualization,Manual feature engineering, Data set description, Automatic feature selection, Model selection, Model training and validation, Model comparison.

We are focusing on developing a system that will predict the salary based on different parameters used in the company/organization in our case from the census bureau and applying the above mentioned methodology phases. Some of the parameters we collected from census data are: Job Type: Private, Self-emp-not-inc, Local_gov.

1.Education: Bachelors, Some-college, 11th, HS-grad.

2.Marital-status: Married-civ-spouse, Divorced, Never-married, Separated.
3.Hrs_of_work per week.
4.Race,Sex,capital_gain and capital loss.
5. Income(predicting variable).

The calculations that will be performed for working of this proposed system to predict the salary with results:

STEP 1:DATA COLLECTION

We start the first one by examining the dataset through the dataset and identifying each and every data type variable and identifying to which datatype they belong to using NOIR datatypes like (nominal,ordinal,interval,ratio) and in adult dataset most of them fall under Nominal(categorical).

STEP 2: DATA CLEANING:

1. Before going to fit the model,Since this data is categorical first you need to factorize the data in the form of levels and then check for any missing values if possible replace them with median,or some information in worst case delete it if it's not needed.

2.After that undergo feature engineering by reducing the factors if any extra or not used which can be achieved through limited factor variables like in case of marital attribute there is divorced and separated both resemble the same thing so try to check for these kind of patterns and reduce the number of factor levels.Since we can't predict the algorithm with different variables having so many levels in my case where there is Country attribute with around 50 levels after doing feature engineering we have narrowed it down to around 5 levels.respectively we done them for all the possible attributes.

STEP 3:DATA VISUALIZATION:

Figure -1: MISSINGMAP

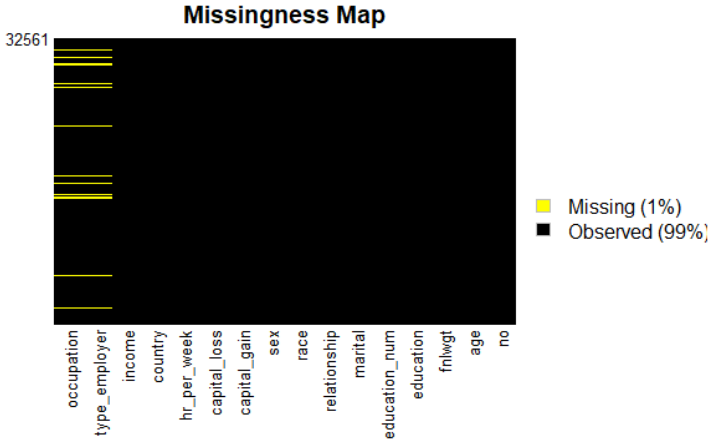


Figure-2:AGE IN FORM OF INCOME

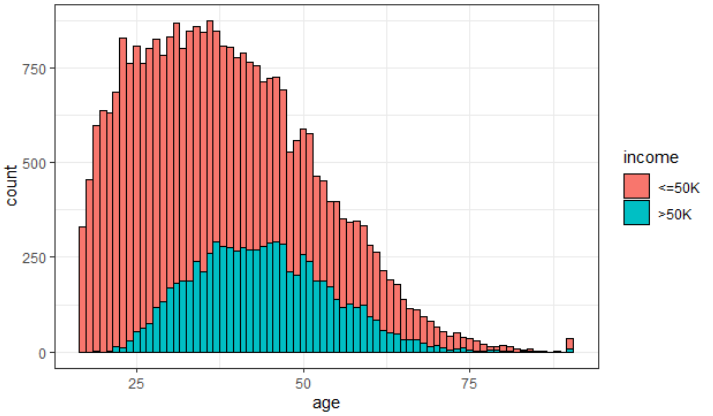


Figure 3:boxplot of age vs education

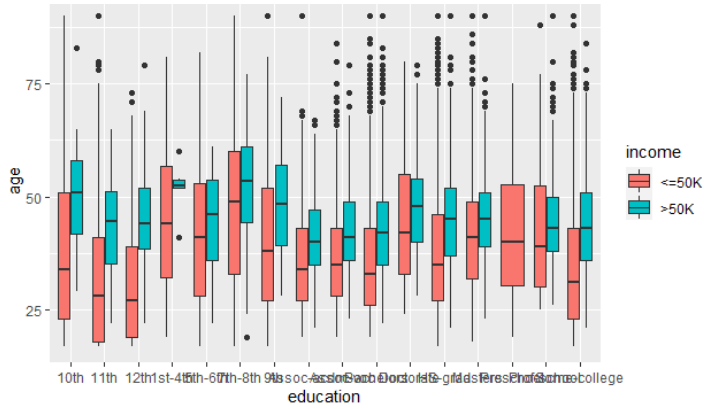
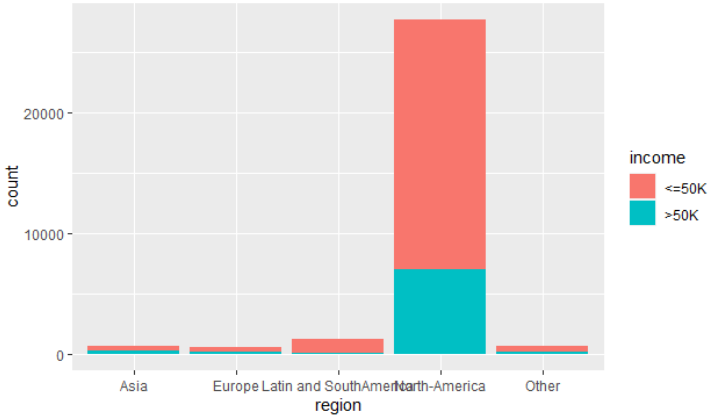


Figure 4:income and region distribution



The importance of data visualization is to understand any patterns which can be identified from the visualization of various variables. We can make some notes about the variables that we may consider for predicting the model accurately without going for best subset selection.

1.The above figure 1 is developed with the dataset having missing values using missmap which is an inbuilt R functionality Which shows whether there are any missing values in it. In our case there are yellow lines which indicate there are missing values in the adult_salary data so we use **na.omit()** to delete all those null values.

2.The figure 2 is a histogram to visualize age with income as colour so we will get an clear idea of how age is affecting income.

3.Figure 3 is a box plot representation of age versus education with income as colouring variable to visualize the variations in the Graph which was created by (age,education) and we can identify that the person with a high age had a higher degree in education.

4.Figure 4 gives you an idea of region and income for which We used bar graph to visualize and we came to know most of the Data is falling under NorthAmerica and most people income is Above 50k in North America.

And below is how we applied the models we discussed earlier and they were explained one by one and all three models will be included in this paper. Step 5 is the building the model process.

Step 4: SPLITTING DATA TO TRAIN AND TEST

1. Now our main challenge is to start splitting the data to get the exact predictions by training it on one type of similar data and applying it on test data which is the other part of dividing pieces.
2. The Amelia package has to be imported to split the data on the basis of splitting data using splitRatio.
3. Which can be done using sample Function and we used 70% data for training and 30% data to check our model.
4. Mean squared error (MSE) and accuracy will be evaluated now along with accuracy to evaluate the baseline model's performance.

Figure -2: Splitting data model

```
# splitting the sample data
sample <- sample.split(adult$income, SplitRatio = 0.70) # SplitRatio = percent

# Training Data
train = subset(adult, sample == TRUE)

# Testing Data
test = subset(adult, sample == FALSE)
```

Step 5:BUILDING THE MODEL

In this project I have developed the models in both R and python language and I used Sql to develop the schema of the dataset and stored the entire csv in it.I have used different machine learning like Linear Regression,Decision Tree,Random Forest since they can be applied on categorical data From these 3 ways we are continuing with applying sklearn train_test_split them on cleaned_data to get accuracy accordingly and apply to it and judge them based on the accuracy.

1.LOGISTIC REGRESSION:

In statistics the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

Now we are using train_test_split to apply to the logistic regression using glm function to perform logistic regression in R and inside it giving the formulae for it and data since I am starting to predict the model here I am gonna use train that we have developed before and use it for building the model and using the test dataset we have kept aside to predict the model.

1. We will be using the glm class from the glm function for this purpose since we are using R. When we are using this function we are gonna give data=train,formulae=’income~.’,and family=binomial(‘logit’) and running it and saving it as model.glm .

```
model.glm<-glm(income~.,family=binomial(link='logit'),data=train)
summary(model.glm)
### More starts menas more significant to predicting the model
new_model<-step(model.glm)
summary(new_model)
test$predicted.income = predict(model.glm, newdata=test, type="response")
test$predicted.income
table(test$income, test$predicted.income > 0.5)
```

2.After creating the model I have used test data to predict the income using predict function with data=test and model name(model.glm) and type=’response’ this is a very important one because we are storing it as another variable so using response is the way to do it.

3. Fit and Transform the variables with test data and then create a Linear Regression model to predict.

4. And we use a confusion matrix to develop the accuracy through the formulae I used in the figure 8 below.

5.I have used step function on model to get the best AIC for different combination of attributes and predicting the income which in my case is best and the value Step: AIC=14122.3 is the best case to choose and the attributes are regage,type_employer,education,marital,rase,sex,hr_per week.

```
> table(test$income, test$predicted.income > 0.5)

       FALSE TRUE
  <=50K  6375  545
  >50K    873 1422
```

Figure -8:transformation confusion matrix

6. Predict the accuracy and precision and recall values.

```
> ## calculating accuracy
> (6375+1422)/(6375+1422+545+873)
[1] 0.8461205
> ## calculating precision
> 6735/(6375+873)
[1] 0.9292219
> ## calculating recall
> 6735/(6375+545)
[1] 0.9732659
>
```

Figure -4: Accuracy,precision,recall values.

This model has very good recall value and I have to check with others to get the final say on which predicting model is the best because we are training the same data with different models to draw a conclusion about it .

Next we are gonna use decision tree and random forest accordingly.

2.DECISION TREE:

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

1. We will use the sklearn library from sklearn. We import decisiontreeclassifier and use the dtree() function to fit the data. This time we are using python to predict the model and we will use a jupyter notebook to run it.

```
model.glm<-glm(income~.,family=binomial(link='logit'),data=train)
summary(model.glm)
### More starts menas more significant to predicting the model
new_model<-step(model.glm)
summary(new_model)
test$predicted.income = predict(model.glm, newdata=test, type="response")
test$predicted.income
table(test$income, test$predicted.income > 0.5)
```

2.In Order to apply the decision tree to the dataset we have to change the categorical data into numeric data by assigning values accordingly and applying the decision tree since it works only on numeric data. Below image shows you how it's done through map function.

```
x={'Husband':28,
'Not-in-family':29,
'Own-child':31,
'Unmarried':32,
'Wife':33,
'Other-relative':34}
df['relationship']=df['relationship'].map(x)
```

Similarly I have applied these to all the other attributes that are available in the form of categorical and changing them accordingly.Here is the model fit.

```
y=df1['income']
X = df1[features]
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X,y)
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30)
```

3.After creating the model I have used test data to predict the income using predict function and here is how its Predicted and compared with other models.

3..RANDOM FOREST:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set.

1. We will use the sklearn library from sklearn. We import decisiontreeclassifier and use the randomForestClassifier() function to fit the data. This time we are using python to predict the model and we will use a jupyter notebook to run it.

2.In Order to apply the random forest to the dataset we have to change the categorical data into numeric data by assigning values accordingly and applying the random forest since it works only on numeric data. Below image shows you how it's done through map function.

```
x={'Husband':28,
'Not-in-family':29,
'Own-child':31,
'Unmarried':32,
'Wife':33,
'Other-relative':34}
df['relationship']=df['relationship'].map(x)
```

Similarly I have applied these to all the other attributes that are available in the form of categorical and changing them accordingly.Here is the model fit.

```
y=df1['income']
X = df1[features]
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X,y)
```

```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier(n_estimators=100)
rfc.fit(X_train, y_train)
```

```
RandomForestClassifier()
```

```
rfc_pred = rfc.predict(X_test)
```

```
print(confusion_matrix(y_test,rfc_pred))
```

```
[[1341  796]
 [ 488 6000]]
```

```
print(classification_report(y_test,rfc_pred))
```

	precision	recall	f1-score	support
N	0.73	0.63	0.68	2137
Y	0.88	0.92	0.90	6488
accuracy			0.85	8625
macro avg	0.81	0.78	0.79	8625
weighted avg	0.85	0.85	0.85	8625

3.After creating the model I have used test data to predict the income using predict function and here is how its Predicted and compared with other models.

1.Fit and Transform the variables with test data and then create a clf.RandomForestClassifier.

2. And we use a confusion matrix to develop the accuracy through the formulae I used in the figure below.

3.Fit and Transform the variables with test data and then create a clf.RandomForestClassifier.

4. And we use a confusion matrix to develop the accuracy through the formulae I used in the figure below.

```
# importing cross_val_score for accuracy with cross validation
from sklearn.model_selection import cross_val_score
cross_val_score(DecisionTreeClassifier(), X, y, cv = 10).mean()
```

```
0.86219646799117
```

Decision tree received better accuracy compared to Logistic regression and Random Forest.

IV CONCLUSIONS

In this paper we proposed an income prediction system by using a logistic regression algorithm with sklearn. For the proper income prediction, we found out the most relevant 5 features. The result of the system is calculated by a suitable algorithm by comparing it with other algorithms in terms of standard scores and curves like the classification accuracy, the ROC curve, the Precision-Recallcurve etc. We compared algorithms only for the basic model with only two attributes. Moreover, we continued with the basic model and found out the most appropriate method to add more attributes and with highest accuracy of 92.3% if applied without region and best accuracy of around 86.22% for decision tree.

In future work, we would like to add a graphical user interface to the system and try to save and reuse trained model.and also use more models accordingly with different target variables and develop a best model.

V ACKNOWLEDGEMENT

I would like to express my sincere gratitude to several individuals and organizations for supporting me throughout my Project. First, I wish to express my sincere gratitude to my Professor Dr.Alla G Webb, for her enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped us tremendously at all times in our research. Without her support and guidance, this project would not have been possible. I could not have imagined having a better supervisor in our study.

VI REFERENCES

[1] Susmita Ray," A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com IT-Con), India, 14th -16th Feb 2019

[2] Sananda Dutta, Airiddha Halder, Kousik Dasgupta," Design of a novel Prediction Engine for predicting suitable salary for a job" 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN).

[3] . Lori Foster Thompson, Phillip W. Braddy, and Karl L. Wuensch. E-recruitment and the benefits of organizational web appeal. Computers in Human Behavior, 24(5):2384 – 2398, 2008. Including the Special Issue: Internet Empowerment.

[4] Phuwadol Viroonluecha, Thongchai Kaewkiriya," Salary Predictor System for Thailand Labour Workforce using Deep Learning" The 18th International Symposium on Communications and Information Technologies (ISCIT 2018)

[5] Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.