

AN IMPLEMENTATION OF SUBSIDY PREDICTION SYSTEM USING MACHINE LEARNING LOGISTICAL REGRESSION ALGORITHM

K. SRINIVAS¹, G. MADHUKAR RAO, K. VENGATESAN, P. SHIVKUMAR TANESH, A. KUMAR,
AND S. YUVARAJ

ABSTRACT. Data mining is the process of extracting new information from the existing data, in which classification plays a major role in different real-time data analytics problems. In our proposed system, we are going to design a classification algorithm for a subsidy delivery system based on the income data set. Generally, subsidy will be assigned based on a few important parameters like income, demographic and few financial parameters. Our proposed system will develop a new classifier system for individuals for delivering the subsidy and also simplify the data system by reducing the number of variables to be considered, without sacrificing too much on accuracy. Such a system will surely help for planning subsidy outlay, monitoring and preventing misuse.

1. INTRODUCTION

Regression analysis is a type of predictive modeling technique, which finds the relationship between the dependent and independent variables. In general, the regression is classified into three types such as linear regression, logistic regression and polynomial regression. In this work, we will discuss how logical regression helps to solve data analysis problems. The logical regression produces the result in binary format, which is used to predict the outcomes of the categorical values. Generally final outcomes will be either 0 or 1, true or false,

¹corresponding author

2010 *Mathematics Subject Classification.* 62M10, 68T09.

Key words and phrases. Logistic Regression, Data Analysis, Classification, Accuracy.

yes or no, high or low. The general mathematical equation of logistical regression is written as follows, where Y is the dependent variable based on X_1, X_2 .

$$Y = C + B_1X_1 + B_2X_2 + \dots$$

Logistic regression is a classification algorithm used to give judgements to a discrete course of action of classes. Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), Y , can take only discrete values for the given set of features (or inputs), X . Contrary to popular belief, logistic regression IS a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as '1'. Just like linear regression, it assumes that the data follows a linear function. For further reference see [1-8].

Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself. Fig. 1 demonstrates the graphical representation of logical regression and linear regression.

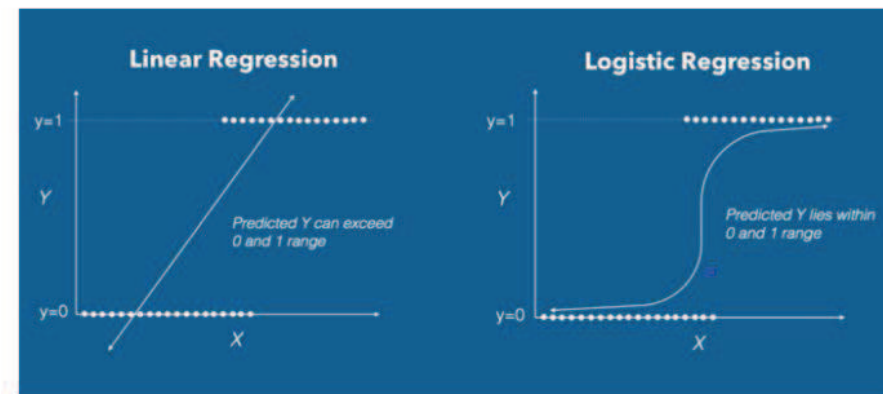


FIGURE 1. Linear Regression VS Logistic Regression Graph

The decision for the value of the threshold value is majorly affected by the values of precision and recall. Ideally, we want both precision and recall to be 1, but this seldom is the case. The decision for the estimation of the breaking point regard is altogether impacted by the estimations of precision and audit. Ideally, we need both precision and survey to be 1.

1.1. Low Precision/High Recall. In applications where we have to decrease the number of fake negatives without in a general sense reducing the number counterfeit positives, we pick a decision worth which has a low estimation of Precision or high estimation of Recall.

1.2. High Precision/Low Recall. In applications where we have to decrease the amount of some positives without basically reducing the number of fake negatives, we pick a decision worth which has a high estimation of Precision or low estimation of Recall.

2. PROPOSED WORKING MODEL

The proposed model working is represented in Fig. 2, which takes input as data set, then we need to identify our main problem statement, problem conceptualization, the suitable methods to solve that problem statement based on that we need to identify. After identification of the problem using classification techniques like Logistical regression, we will calculate confusion matrix and accuracy score to verify our results with training and test data set.

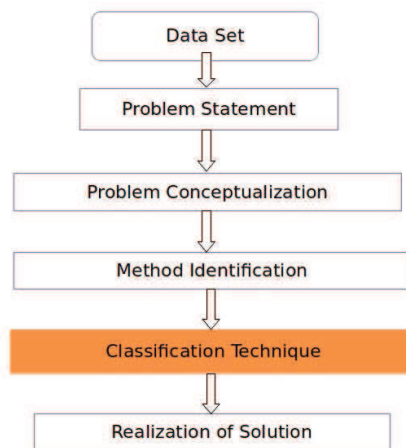
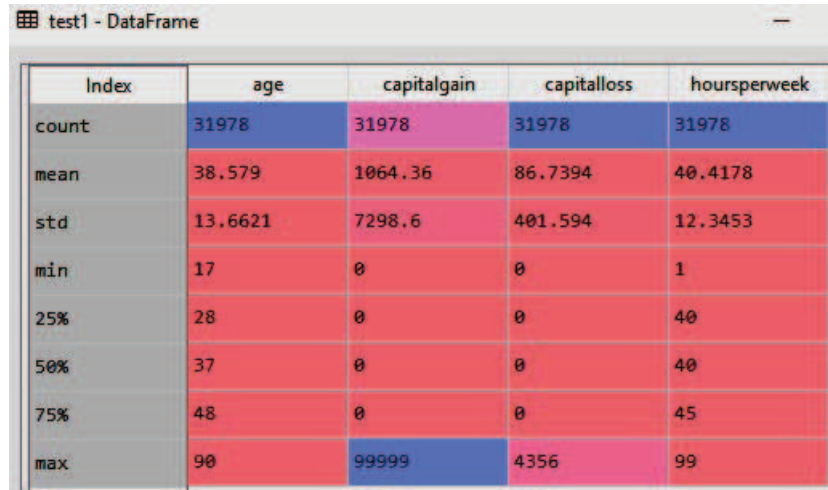


FIGURE 2. Classification Model to find the income Bracket

3. RESULTS AND DISCUSSIONS

In this work, we are going to consider sample data set 'income.csv' with 31973 rows and 13 columns with 4 numerical data types and 9 categorical data types.

This data consists of columns like Job Type, EdType, marital status, occupation, relationship, race, gender, capital gain, capital loss, hoursperweek, native county and Sal Stat. Our main aim to develop an income classifier with a reduced number of variables from the considered income data set. The proposed system uses the data set, which is supplied to the classification algorithm, from which we develop a model based on the training data. Those results are validated using test data. Finally, the least required variable is used to build the income classification model. The first phases of our proposed work is identifying the missing values, from the data set and apply the proper pre-processing techniques to clean the missing values. Identifying the most influencing variable from the data set is another important task with respect to salary status variable, which can be achieved using statistical techniques like correlation, chi-square test. Visualization is demonstrated by box plot and scatter plot. Fig. 3 shows the basic statistical information about income data set like mean, standard deviation, minimum, maximum values of corresponding attributes. Fig. 4 will represent the correlation between the numerical attributes like age, capital gain, capital loss and hours per work from the income data set.



Index	age	capitalgain	capitalloss	hoursperweek
count	31978	31978	31978	31978
mean	38.579	1064.36	86.7394	40.4178
std	13.6621	7298.6	401.594	12.3453
min	17	0	0	1
25%	28	0	0	40
50%	37	0	0	40
75%	48	0	0	45
max	90	99999	4356	99

FIGURE 3. Statistical information about income data set

Fig. 5 represents the frequency distribution of SalStat, which is visualized into two categories: salary greater than and less than 50,000. Many of the attributes are less than 50,000. Fig. 6 represents the age attribute visualized into a histogram, into 10 bins from 0 to 90. The graph demonstrates the maximum ages

Index	age	capitalgain	capitalloss	hoursperweek
age	1	0.0774899	0.057282	0.0682992
capitalgain	0.0774899	1	-0.0314986	0.0791118
capitalloss	0.057282	-0.0314986	1	0.0531065
hoursperweek	0.0682992	0.0791118	0.0531065	1

FIGURE 4. Correlation between the attributes

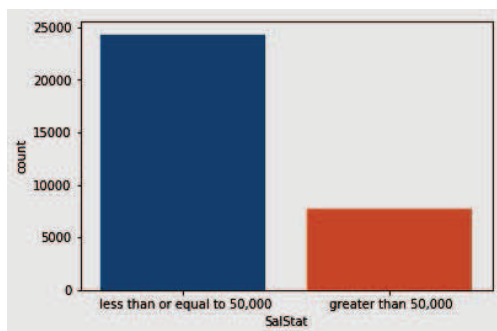


FIGURE 5. Frequency Distribution of SalStat

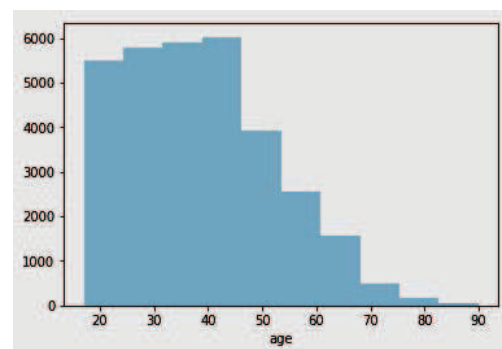


FIGURE 6. Histogram on Age

are from 40 to 50 years. Fig. 7 represents the bar plot between occupation vs SalStat, maximum occupations are adm-clerical, craft-repair and other services.

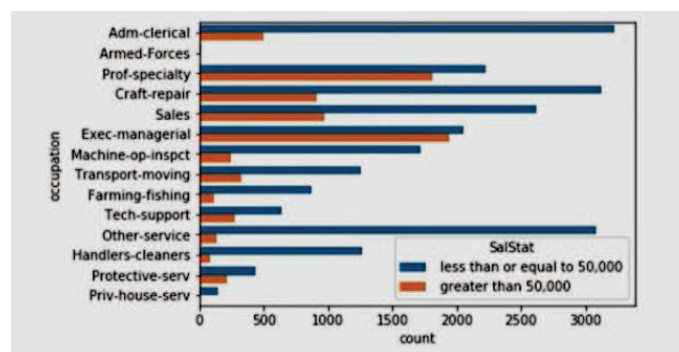


FIGURE 7. Occupation Vs SalStat

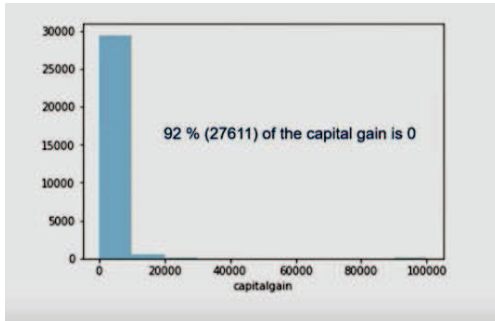


FIGURE 8. Capital Gain vs SalStat

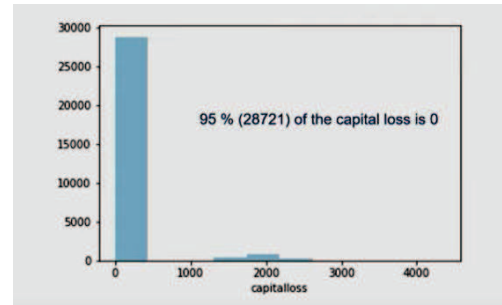


FIGURE 9. Capital Loss vs SalStat

Fig. 8 shows the SalStat vs Capitalgain, 27611 persons have zero relationships with SalStat from income data set. Fig. 9 shows SalStat vs Capitalloss, 28721 persons have zero relationships with SalStat from income data set.

The actual implementation is tested using logistic regression and the KNN algorithm, in which the income data set contains SalStat, which are categorical variables. As per the analytics technique, the categorical variable would not classify directly, so those values will be converted into a number of types using Map() function. The SalStat attribute is converted into numerical types. The data set is tested using logistic regression and the KNN algorithm, whose accuracy values nearly to 84.02%.

Logistic regression is a machine learning classification algorithm that is used to predict the probability of the categorical dependent variable. Using logistic regression, we will build a classification model based on the income data set. First, the SalStat variable is modified into 0 or 1. Fig. 10 represents the SalStat column converted into categorical and numerical attributes values either 0 or 1 by using the map function. Fig. 11 represents the total income that is converted into dummy values based on the SalStat column attribute.

4. ACCURACY CALCULATION

Fig. 12 shows the logistic regression, which is calculated based on the confusion matrix and accuracy score function values for the income data set. The final accuracy value is 0.85. These values are tested using the training data set.

Index	capitalgain	capitalloss	hoursperweek	nativecountry	SalStat
0	0	0	28	United-States	0
1	0	0	40	United-States	0
2	0	0	40	United-States	1
3	0	0	40	Mexico	0
4	0	0	35	United-States	0

FIGURE 10. Conversion of categorical variables into numerical for SalStat Variables

Index	age	capitalgain	capitalloss	hoursperweek	SalStat
0	45	0	0	28	0
1	24	0	0	40	0
2	44	0	0	40	1
3	27	0	0	40	0
4	20	0	0	35	0

FIGURE 11. Income data set converted into dummy values

```

In [20]: logistic.intercept_
Out[20]: array([-3.56758886])

In [21]: prediction=logistic.predict(test_x)

In [22]: confusion_matrix=confusion_matrix(test_y,prediction)

In [23]: accuracy_score=accuracy_score(test_y,prediction)

In [24]: print(accuracy_score)
0.8528246820929748

```

FIGURE 12. Accuracy Calculation using logistic regression

5. CONCLUSION

Data analytic provides various techniques to investigate the various real-time data set applications, and to predict the future analysis of related domain or area. In this work we have proposed data mining logistical regression which will predict the subsidy of the customers based on the salary attribute. It generates the confusion matrix based on the training data set and the performance of the proposed system is tested using an accuracy score function. Our system selects explicit attributes from the data set to assemble the logistic regression system which produces an accuracy of about 85.02 %.

REFERENCES

- [1] R. BUSS: *The New Basics: Today's Employers Want "Three Rs" and So Much More*, Vocational Educational Journal, **67**(5) (2018), 24-25.
- [2] L. HARVEY: *New Realities: The Relationship Between Higher Education And Employment*, Territory Education and Management, **6** (2012), 3-17.
- [3] K. VENGATESAN, E. S. KUMAR, S. YUVARAJ, P. S. TANESH, A. KUMAR: *An Approach for Remove Missing Values in Numerical and Categorical Values Using Two Way Table Marginal Joint Probability*, International Journal of Advanced Science and Technology, **29**(5) (2020), 2745 - 2756.
- [4] K. VENGATESAN, S. YUVARAJ, S. SAMEE, P. S. TANESH, A. KUMAR: *Predition of the Petrol Consumptions by using Data Mining Decision Classifier Classification Algorithm*, TEST Engineering and Management, **45** (2020), 22206 - 22212.
- [5] R. PUNNOOSE, P. AJIT: *Prediction of Employee Turnover in Organizations using Machine Learning Algorithms—A Case for Extreme Gradient Boosting*, IJARAI, **5**(9) (2016), 112–118.
- [6] S. KAUR, S. JINDAL: *A Survey on Machine Learning Algorithms*, 2nd Edition, Wiley, 2016.
- [7] S. N. SIVANANDAM, S. N. DEEPA: *Principles of Soft Computing*, 2nd Edition, Wiley, 2018.
- [8] V. N. KALBANDE, C. C. HANDA: *Developing A Model To Predict Employability Of Engineering Students In Campus Placement For IT Sector*, IJAREST, **2**(6) (2015), 33–44.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
KONERU LAKSMAIAH EDUCATION FOUNDATION
HYDERABAD,500075,INDIA
E-mail address: srirecw9@klh.edu.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
KONERU LAKSMAIAH EDUCATION FOUNDATION
HYDERABAD,500075,INDIA
E-mail address: madhusw511@gmail.com

SANJIVANI COLLEGE OF ENGINEERING
SAVITRIBAI PHULE UNIVERSITY
MH,INDIA.
E-mail address: vengicse2005@gmail.com

SANJIVANI COLLEGE OF ENGINEERING
SAVITRIBAI PHULE UNIVERSITY
MH,INDIA.
E-mail address: punjabishiv@gmail.com

SCHOOL OF CS AND IT
JAIN UNIVERSITY
BANGALORE, INDIA
E-mail address: abhishek.maacindia@gmail.com

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPT OF ECE
CHENNAI, INDIA
E-mail address: yuvasivasanthi@gmail.com