IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11354

Prediction through machine learning on the dependence of job prospects in the Afro-American community on proficiency in English

Animesh Samanta[1], Akash Chowdhury[2], Dip Das[3], Arup Kumar Dey[4], Mrs. Sulekha Das[5]

[1] UG-Computer Science and Engineering, Techno Engineering College Banipur,

[2] UG-Computer Science and Engineering, Techno Engineering College Banipur,

[3] UG-Information Technology, Techno Engineering College Banipur,

[4] UG-Information Technology, Techno Engineering College Banipur,

[5] Assistant Professor, Techno Engineering College Banipur

Abstract: In the international business sphere, English has become the lingua-franca of the business world irrespective of geographical, social, political, or religious differences. With jobs becoming more and more global, English as a global language has gained importance as a medium of communication, both at the international and intra-national levels. In the professional world, communication skills are very crucial. Being proficient in English means being able to communicate clearly and effectively. Enhanced communication skills in English can help attain better/ advanced education and consequently aid in availing better job opportunities in the future. In this paper, we explore two crucial aspects of the assimilation experience of colored Afro-Americans. It explores the determinants of their English language (speaking) fluency and the key role such skills play in their occupational success. We find that in a particular population enhanced level of fluency in English results in brighter job prospects. Advance education and better jobs help in keeping the population largely away from involvement in unlawful activities. Data has been analyzed through Multiple Regression Analysis (MRA). The proposed model is tested on the "Communities and Crime Data Set" from the UCI Machine Learning Repository: which is available at https://archive.ics.uci.edu/ml/datasets/communities+and+crime.

1. INTRODUCTION

The English language is essential in our present society and worldwide also to expose our views throughout the world. In education, research, business, agriculture, and every professional sector without English, we can't develop ourselves and also our country. Based on the survey in the Afro-American community about the speaking ability in English in a particular population who are 16 and above working as an employee and those are employed in professional services. Many times those who work do not know English well but knowing English is very important in a job, especially in a professional service. It is most difficult to do most of the work without English. Besides you need to speak English fluently, if someone can't speak English very well, he or she can't explain anything and not be able to express his or her talent to anyone else and face a lot of problems at work. Through multiple regression using machine learning, we have predicted the speaking ability in the above-mentioned criteria.

Machine learning can simply be defined as using data instead of logic to perform tasks by a machine. We use data to train the machine, as in, tell it what it has to do and then test the trained model on different tasks to see whether the training has been successful or not. When it comes to data mining, the term classification plays an important role as it assigns class values to new instances found during data mining [8].

Multiple regression is an applied math technique that will be accustomed analyze the link between one dependent variable and several other independent variables. The target of multiple regression analysis is to use the independent variables whose worths square measure is acknowledged to predict the worth of the only dependent value.

Cross-validation (CV) is a popular strategy for algorithm selection. The main idea behind CV is to split data, once or several times, for estimating the risk of each algorithm: Part of the data (the training sample) is used for training each algorithm, and the remaining part (the validation sample) is used for estimating the risk of the algorithm. Then, CV selects the algorithm with the smallest estimated risk [4]. Besides the common case, where a general measure of the reliability and accuracy of a system is needed, evaluation often becomes necessary to choose the best one out of different methods and/or parameter sets[6].

DOI: 10.17148/IJARCCE.2022.11354

Statistical analysis is widely used in all aspects such as in science, medicine, crime, English literature, employed in professional sources, and also in social sciences. There are many methods in statistics and one of them is regression. There are six types of linear regression analyses which are simple linear regression, multiple linear regression, logistic regression,

ordinal regression, multinominal regression, and discriminant analysis [1]. Multiple linear regression was selected to build a model of prediction on the dependence of job prospects in the Afro-American community on proficiency in English. One method that is categorized in the stepwise-type procedures is stepwise regression also used in this paper. The main objective of this paper is to select the suitably controlled variables in the forecast for the prediction of the dependence of job prospects in the Afro-American community on proficiency in English [2].

2. LITERATURE REVIEW

Multiple regression is an applied math technique, the target of multiple regression analysis is to use the independent variables whose worths square measure acknowledged to predict the worth of the only dependent value. The main aim of this project is the prediction of the dependence of job prospects in the Afro-American community on proficiency in English. Mr. M. S. BARTLETT had done on “FURTHER ASPECTS OF THE THEORY OF MULTIPLE REGRESSION”.

Intan Martina Md Ghani, Sabri Ahmad (2010) had done research to Forecast Fish landings.

Isık Yilmaz and Oguz Kaynar (2011) had done research prediction of the swell potential of clayey soils using multiple linear regression.

We have gathered some specific ideas about machine learning and multiple linear regression. So we were interested very in doing a project based on it. So we had collected some real-life data on the “Communities and Crime Data Set” from the UCI, which is available at https://archive.ics.uci.edu/ml/datasets/communities+and+crime. and try to predict speaking ability in English fluently using independent fields PctEmploy Population and PctEmplProfServ.

3. METHODOLOGY

In this paper, data were taken from UCI Machine Learning Repository. Here we work in the following field…

Table.1. Data Field

Attributes	**Description**	**Mean**
PctEmploy	percentage of people 16 and over who are employed	0.501
Population	population for community	0.057
PctSpeakEnglOnly	percent of people who speak only English.	0.785
PctEmplProfServ	percentage of people 16 and over who are employed in professional services	0.440

4. RESEARCH METHOD

Multiple simple regression is the strategy of statistics in regression that's familiar to analyzing the link between one response variable (dependent variable) with 2 or additional controlled variables (independent variables). This methodology was selected for this analysis as a result there have been quite controlled variables. during this analysis, the response variable is Communication In English Only(Y) Employee (X_1), Population (X_2), Employee in Professional Services (X_3) are controlled variables.

DOI: 10.17148/IJARCCE.2022.11354

4.1 Accuracy of difference between Actual data and Calculated data

In this research, the hypotheses that were used*:*
H0: $b_1=b_2=b_3=b_4=0$
Ha: At least one of the b_1, b_2, b_3, and b_4 does not equal 0 which says that
H0: None of the controlled variables X_1, X_2, X_3, and X_4 is significantly related to Y
Ha: At least one of the controlled variables X_1, X_2, X_3, and X_4 is significantly related to Y The model of multiple regression can be represented as:

$$\mathbf{Y = a + b_1X_1 + b_2X_2 + \ldots \ldots \ldots \ldots . + b_nX_n}$$

where
y= Dependent variable (Speak English only)
a=Constant variable
b_1=Coefficient of the first control variable,
b_2=Coefficient of the second control variable,
b_3=Coefficient of the third control variable,
x_1=controlled variable(employee)
x_2=controlled variable(population)
x_3=controlled variable (employee in professional services)

4.2 Confusion-Matrix

After finding the accuracy of the difference between actual data and calculated data we did the Confusion Matrix. In this confusion matrix it can be seen that,[2] we find the **TP** – which stands for '**TRUE POSITIVE'** means the accuracy of classified positive data, **TN** – which stands for '**TRUE NEGATIVE'** means the accuracy of classified negative data, **FP** – which stands for '**FALSE POSITIVE'**, means which remark that actual value is negative but predicted data is positive, **FN** – which stands for '**FALSE NEGATIVE'** means that actual data and the predicted data both are negative and append the TP, TN, FP, FN value in 2*2 matrix(mat1). After that, we find the accuracy, sensitivity, precision, recall, and specificity. This matrix contains all the raw information about the predictions done by a classification model on a given data set.[3]

4.3 Cross-Validation

After finding the accuracy of the difference between actual data and calculated data we did cross-validation. In this cross-validation process first, we divide the whole list into 10 sub-list and then we find the accuracy of 10 sub-list elements we also find the Confusion Matrix of each Sub-list and we find the accuracy, and sensitivity, precision, recall, and specificity.

ACCURACY: It's the ratio of the correctly labeled subjects to the whole pool of subjects.

Accuracy is intuitional.
PRECISION: Precision is the ratio of the correctly +ve labeled by our program to all +ve labeled.
RECALL: Recall means out of the total positive, what percentage are predicted positive.
SPECIFICITY: Specificity is calculated as the number of correct negative predictions divided by the total number of negatives.

- **ACCURACY= (TP+TN/ TP+TN+FP+FN)* 100**
- **PRECISION = (TP/FP+TP)*100**
- **RECALL= (TP/FN+TP)*100**
- **SPECIFICITY = (TN/TN+FP)* 100**

DOI: 10.17148/IJARCCE.2022.11354

4.4 Flow Chart

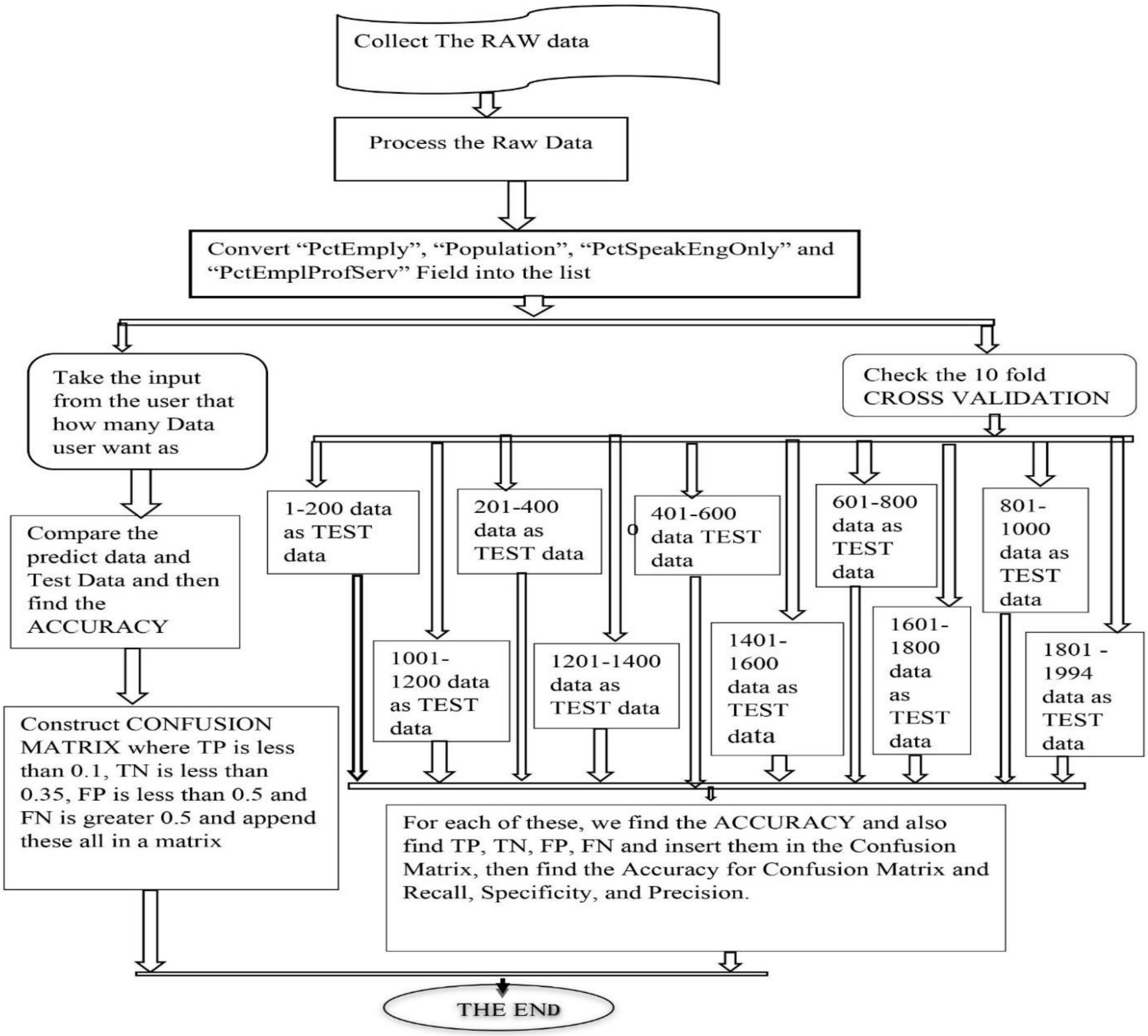

Collect The RAW data
Process the Raw Data
Convert "PctEmply", "Population", "PctSpeakEngOnly" and "PctEmplProfServ" Field into the list
Take the input from the user that how many Data user want as
Check the 10 fold CROSS VALIDATION
1-200 data as TEST data
201-400 data as TEST data
401-600 data TEST data
601-800 data as TEST data
801-1000 data as TEST data
Compare the predict data and Test Data and then find the ACCURACY
1001-1200 data as TEST data
1201-1400 data as TEST data
1401-1600 data as TEST data
1601-1800 data as TEST data
1801 - 1994 data as TEST data
Construct CONFUSION MATRIX where TP is less than 0.1, TN is less than 0.35, FP is less than 0.5 and FN is greater 0.5 and append these all in a matrix
For each of these, we find the ACCURACY and also find TP, TN, FP, FN and insert them in the Confusion Matrix, then find the Accuracy for Confusion Matrix and Recall, Specificity, and Precision.
THE END

Fig 1

5. RESULT

Table.2. Accuracy of difference between Actual data and Calculated data

Accuracy of 90%Data as Training Data or(0.9)	**84.92**
Accuracy of 80%Data as Training Data or(0.8)	**85.17**
Accuracy of 75%Data as Training Data or(0.75)	**85.74**
Accuracy of 66%Data as Training Data or(0.66)	**84.63**

Table.3. Confusion Matrix & Corresponding Result

For 90% of Data	**For 80% of Data**
Confusion Matrix: 152 30 5 12	**Confusion Matrix:** 316 51 10 21
Accuracy: 91.46	**Accuracy:** 93.77

Precision: 96.82 **Recall:** 92.68 **Specificity:** 85.71	**Precision:** 96.93 **Recall:** 93.77 **Specificity:** 83.61
For 66% of Data **Confusion Matrix:** 524 95 16 42 **Accuracy:** 91.43 **Precision:** 97.04 **Recall:** 92.58 **Specificity:** 85.59	**For 50% Data** **Confusion Matrix:** 778 129 32 57 **Accuracy:** 91.06 **Precision:** 96.05 **Recall:** 93.17 **Specificity:** 83.57

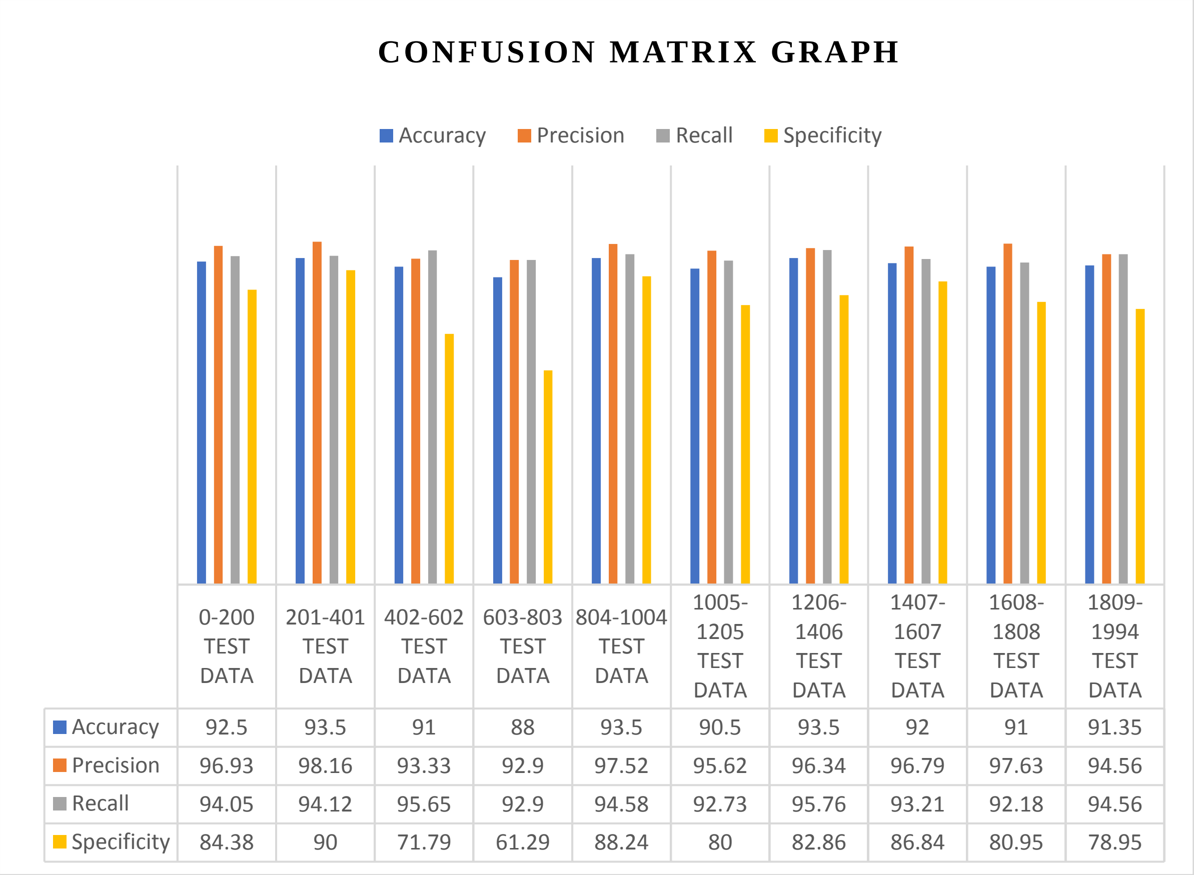
Table.4. For 10-fold cross-validation Accuracy

TEST CASE	ACCURACY RATE (%)
1	86.0
2	89.5
3	84.5
4	84.5
5	87.0
6	85.5
7	87.0
8	83.5
9	88.5
10	83.24

Table.5. For 10-fold cross-validation Results

0-200 Test Data Confusion Matrix: 158 27 5 10 Accuracy: 92.5 Precision: 96.93 Recall: 94.05 Specificity: 84.38	**201-401 Test Data** Confusion Matrix: 160 27 3 10 Accuracy: 93.5 Precision: 98.16 Recall: 94.12 Specificity: 90.0
402-602 Test Data Confusion Matrix: 154 28 11 7 Accuracy: 91.0 Precision: 93.33 Recall: 95.65 Specificity:71.79	**603-803 Test Data** Confusion Matrix: 157 19 12 12 Accuracy: 88.0 Precision: 92.9 Recall: 92.9 Specificity: 61.29
804-1004 Test Data Confusion Matrix: 157 30 4 9 Accuracy: 93.5 Precision: 97.52 Recall: 94.58 Specificity: 88.24	**1005-1205 Test Data** Confusion Matrix: 153 28 7 12 Accuracy: 90.5 Precision: 95.62 Recall: 92.73 Specificity: 80.0

1206-1406 Test Data	1407-1607 Test Data
Confusion Matrix: 158 29 6 7 Accuracy: 93.5 Precision: 96.34 Recall: 95.76 Specificity: 82.86	Confusion Matrix: 151 33 5 11 Accuracy: 92.0 Precision: 96.79 Recall: 93.21 Specificity: 86.84
1608-1808 Test Data	**1809-1994 Test Data**
Confusion Matrix: 165 17 4 14 Accuracy: 91.0 Precision: 97.63 Recall: 92.18 Specificity: 80.95	Confusion Matrix: 139 30 8 8 Accuracy: 91.35 Precision: 94.56 Recall: 94.56 Specificity: 78.95



	0-200 TEST DATA	201-401 TEST DATA	402-602 TEST DATA	603-803 TEST DATA	804-1004 TEST DATA	1005-1205 TEST DATA	1206-1406 TEST DATA	1407-1607 TEST DATA	1608-1808 TEST DATA	1809-1994 TEST DATA
Accuracy	92.5	93.5	91	88	93.5	90.5	93.5	92	91	91.35
Precision	96.93	98.16	93.33	92.9	97.52	95.62	96.34	96.79	97.63	94.56
Recall	94.05	94.12	95.65	92.9	94.58	92.73	95.76	93.21	92.18	94.56
Specificity	84.38	90	71.79	61.29	88.24	80	82.86	86.84	80.95	78.95

Fig.2. 10-fold cross-validation Confusion Matrix Graph

CONCLUSIONS

This paper uses multiple regressions (MLR) to predict the crime level. We have collected the data from UCI Machine Learning Repository based on that we made a relationship between the dependent variable and the independent variable after that we perform Confusion Matrix where we compare the actual target values with those predicted by the machine learning model. After checking the Confusion Matrix, we move to the Cross Validation where we find the accuracy of 10 sub-list elements and we also find the Confusion Matrix of each Sub-list. we predict the accuracy as well as sensitivity, precision, recall, and specificity for user choice test data and the 10 sub-list. This type of project may help in the future to find any kind of prediction from any data field.

DOI: 10.17148/IJARCCE.2022.11354

REFERENCES

1. Ghani, I. M. M., & Ahmad, S. (2010). Stepwise multiple regression method to forecast fish landing. *Procedia-Social and Behavioral Sciences*, *8*, 549-554.Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
2. Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *MAICS*, *710*, 120-127.
3. Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3), 429-450.
4. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40-79.
5. Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95-112.
6. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192-213.
7. Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In *data democracy* (pp. 83-106). Academic Press.
8. Yadav, S., & Shukla, S. (2016, February). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International conference on advanced computing (IACC)* (pp. 78-83). IEEE.
9. Valente, G., Castellanos, A. L., Hausfeld, L., De Martino, F., & Formisano, E. (2021). Cross-validation and permutations in MVPA: Validity of permutation strategies and power of cross-validation schemes. *NeuroImage*, *238*, 118145.
10. Groening, C., Mittal, V., & "Anthea" Zhang, Y. (2016). Cross-validation of customer and employee signals and firm valuation. *Journal of Marketing Research*, *53*(1), 61-76.
11. Wainer, J., & Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, *182*, 115222.
12. Zhang, D., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, *59*(2), 895-907.
13. Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, *180*, 68-77.
14. Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. (2021). An Enhanced Random Forest Model for Detecting Effects on Organs after Recovering from Dengue. *methods*, *8*(8).
15. Chaudhuri, A. K., Banerjee, D. K., & Das, A. (2021). A Dataset Centric Feature Selection and Stacked Model to Detect Breast Cancer. *International Journal of Intelligent Systems and Applications (IJISA)*, *13*(4), 24-37.