Income Prediction

Name:	**Vimaleswar A**
Registration No./Roll No.:	20305
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	January 12, 2023
Date of Submission:	April 16, 2023

1 Introduction

We are provided with trained labelled dataset X(matrix of 43957, 15) where X is (X_i, y_i), $\forall i = 1, 2, \ldots n$, (Here, n = 43957) be the instances in a training data set, where $X_i = [x_{i1}, x_{i2}, x_{i3}, \ldots x_{im}] \in \mathbb{R}^m$ be a m (Here, m = 15) dimensional feature vector and n be the number of instances, also called as data points. Here $x_{ij}, \forall j = 1, 2, \ldots m$ are m individual features and y_i be the target variable of the data point X_i, $\forall i$. Here, the target variable is discrete. Given predictive features like education, employment status , marital status etc., the objective is to predict, if the salary is greater than 50 k or not. In target column, 1 is denoted for salary greater than 50 k and 0 is denoted for salary lesser than 50k. In the given data, we have both categorical features and numerical feature. Our task is to build a classifier ,i.e., a mathematical model and train its parameters through the given labelled training data set and thereby producing an infered function to a new test point to the most suitable discrete class.

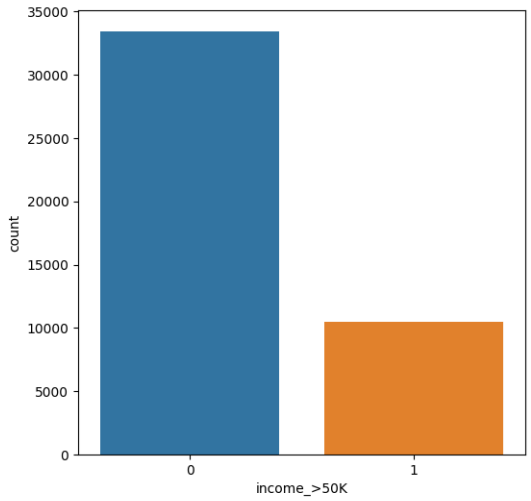

Figure 1: Overview of Class Labels

2 Methods

We have divided the task into parts

- Data Pre-Processing

- Feature Selection
- Model Training
- Testing the model

2.1 Data Pre-Processing

First, we have imported all the necessary python libraries for our ML project. Then, we loaded the train.csv file data set to pandas DataFrame. We then load the train class levels.csv file data set. Next, we have merged the two data sets and name it as income data set. As many machine learning algorithms do not support missing or null values, we have handled the null values in mainly two ways:
i) Deleting rows with missing values (naming it has income dataset1).
ii) Replacing null values with mean/median for numerical columns and mode for categorical columns (naming it has income dataset2).
Next, we work on label encoding. Since, computer can understand only numerical data and not text data, we convert all the categorical columns to numerical columns for both the income datasets. Next we are scaling the features as they are varying in a large range. Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. Here, we use Standardization method.

2.2 Splitting technique

We splitted our given train data set into two different data set in the following proportion: 80% - 20% : one to train our model (training set) and the other to test our model against the data points with a (validation set) with a random state of 30 so as to make the complete use of our data. The stratify parameter will preserve the proportion of target/class label as in original data set, in both the training and validation data sets. We want to preserve the data set proportions for better prediction and reproducibility of results. After that, we have trained the data with different models to check the accuracy and f-score for both the income datasets. After running all the models we found that among all the models **Random Forest Classifier** produces the highest accuracy in **income dataset 2**.

2.3 Feature Selection and Model Training

So, with Income dataset 2 as primary dataset we are plotting correlation matrix. For this, we have used the Pearson Correlation method. Based on theoritical knowledge and from Correlation matrix we are removing the unwanted features present in the dataset which are nearly independent on target variables. After that we are training our dataset with the classifiers which we studied during the coursework and choose the relevant ones which best suit our classification problem. We have used the following classifiers for our problem: A) Decision tree Classifier(DTC) B) Random Forest Classifier(RFC) C) k-Nearest Neighbour Classifier(KNN) E) Logistic regression(LR) F) Guassian Naive Bayes(GNB).
During the evaluation of the model and choosing the set of hyperparameter values we have used f1 as the scoring parameter in the Grid Search Cross validation. It is a better evaluating metric than accuracy as we have discussed in the class from the example of classification problem of Malignant and Benign Tumour. We have given cv as 10 means that the cross validation involves splitting the dataset into 10 folds and train each folds and evaluate per fold. The mean accuracy of the folds along with best hyperparameter is print as the result. We have found that the data set when performed feature selection and scaling(StandardScalar()) has shown the best metric statistics in Random Forest Classifier, Decision Tree Classifier, kNN.

3 Evaluation Criteria

Precision, Recall, F-measure and Accuracy are evaluation metrics commonly used in classification tasks to measure the performance of a predictive model.

Classifier	Accuracy	Precision	Recall	F-measure
RFC	0.85	0.81	0.76	0.78
LR	0.82	0.78	0.70	0.72
DTC	0.85	0.82	0.74	0.76
kNN	0.84	0.79	0.76	0.77
GNB	0.80	0.75	0.63	0.65

Table 1: Performance Of Different Classifiers Using all terms

Predicted Class

	0	1
0	6276	412
1	869	1235

RFC

Predicted Class

	0	1
0	6297	391
1	1155	949

LR

Predicted Class

	0	1
0	6370	318
1	1009	1095

DTC

Predicted Class

	0	1
0	6174	514
1	857	1247

k-NN

Table 2: Confusion Matrices using All Terms

Precision: Precision is the fraction of true positive predictions among all the positive predictions made by the model. It is a measure of the model's ability to avoid false positives. The formula for precision is

Precision = True Positives / (True Positives + False Positives)

Recall: Recall is the fraction of true positive predictions among all the actual positive instances in the data. It is a measure of the model's ability to avoid false negatives. The formula for recall is

Recall = True Positives / (True Positives + False Negatives)

F-measure: The F-measure is the harmonic mean of precision and recall. It is used to balance the importance of precision and recall when evaluating a model's performance. The formula for F-measure is

F-measure = 2 * (Precision * Recall) / (Precision + Recall)

4 Analysis of Results

Since, during the first phase we got that income dataset 2 produces the highest f1 score and accuracy. So, we worked only with the income dataset 2. Then, we performed Hyperparameter tuning using the GridSearch CV technique for all the models where we used **f1** as the scoring criteria to find out the best score and accuracy.

We found that Random Forest Classifier produces the highest best score and accuracy i.e. **67%** and **85%** respectively (with best hyperparameters criterion='gini', max features='sqrt', min samples split=10, n estimators=100, bootstrap=True, max depth=None), Decision Tree Classifier has accuracy of **84.95%** (with best hyperparameter criterion='gini', max depth=10, min samples leaf=3, min samples split=2), kNN has accurqacy of **84.40%** and best score as **64%** (with best hyperparameters metric='manhattan', n neighbors=9, weights='uniform'). We have created different tables Table

1(Performance of different classifiers), Table 2(Confusion Matrices of different classifiers), to show the experimental results.

5 Discussions and Conclusion

After the completion of this project, we can conclude that the methods used in this project are quite intermediate. We have used simple yet rigorous machine learning algorithms. Our data set contains approximately 3000 null values which gave us a good challenge as the presence of null values doesn't allow us to run the model smoothly. So, solving the issue of null values proved to be time-consuming and complex for us.
Overall, we can say that if the data set could have been improved with fewer or no null values and correlation between the features and target variable are high, we would have more accurately predicted whether the income of a person is greater than 50K dollars or not and would have got high best score based on f-score.
On discussing our future plans regarding the project, we will say that as the world is advancing more and more with the introduction of Artificial Intelligence we're thinking to explore more ensembler classifiers like XGM and also deep learning models like ANN(Artificial Neural Network) or Multilayer Perceptron algorithmwe and will also try to improve more with different hyperparameter tuning models using better parameters as there is always better waiting for next.

6 References

Class notes(DSML Spring 2023) **Dr.Tanmay Basu** .
Machine Learning by **Tom Mitchell**
https://scikit-learn.org/
https://scikit-learn.org/stable/supervised$_{l}earning.htmlsupervised - learning$
https://towardsdatascience.com/
https://www.kaggle.com/

7 Github Link

Vimaleswar