# Income Level Prediction

Rachaell Nihalaani
*College of Engineering*
*University of Utah*
u1417978

*Abstract* - **The Income Level Prediction project is a classification problem with the prediction task to determine whether a person makes over 50K a year given the census information. This project has been developed for CS5350/6350 Machine Learning in Fall 2022 at the University of Utah.**

***Index Terms - Machine Learning, Binary Classification,***

## I. PROBLEM DEFINITION AND MOTIVATION

The chosen problem is Income Level Prediction. The goal of this project is to build a machine learning model that can predict an individual's income level based on various demographic and employment characteristics. This is an interesting project as accurate income level prediction can have a variety of applications, including targeted marketing and credit risk assessment. We used machine learning to solve this as classification algorithms used in machine learning utilize input training data for the purpose of predicting the likelihood or probability that the data that follows will fall into one of the predetermined categories.

## II. DATASET DESCRIPTION

This dataset was extracted by by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the specific conditions. There are 14 attributes, including continuous, categorical and integer types and some attributes have missing values, recorded as question marks. The attributes are described in Table 1. The target label ‘income > 50K’ is a binary value [0,1] representing individuals with an annual income greater than or less than $50,000.

Table 1 Data Attributes

| Attribute | Description | Attribute | Description |
|---|---|---|---|
| age | continuous | workclass | categorical |
| fnlwgt | continuous | education | categorical |
| education-num | continuous | marital-status | categorical |
| capital-gain | continuous | occupation | categorical |
| capital-loss | continuous | relationship | categorical |
| hours-per-week | continuous | race | categorical |

| | | | |
|---|---|---|---|
| native-country | categorical | sex | categorical |

We mentioned above that some attributes have missing values, to tackle this, some data preprocessing is required to enhance the prediction capabilities of the models before building it. So, missing values were imputed with the median value for each feature. We also performed feature engineering and used ordinal encoders on categorical features.

## III. PROPOSED SOLUTION

A variety of machine learning algorithms were evaluated for this task, and are listed and explained in brief below.

### A. Algorithms

#### 1. Logistic Regression

This is a machine learning algorithm that is primarily used for predictive analysis and is built on the probability concept. Logistic Regression is a go-to method for binary classification problems. It is a linear regression model which does not use a linear function, and instead makes use of the sigmoid or logistic function, which is a cost function of higher complexity. Equation (1) shows the definition of the logistic sigmoid function. .

$$f(x) = \frac{1}{1+e^{-x}} \qquad (1)$$

Sigmoid functions cannot be represented by linear functions as they can take values less than 0 or greater than 1, and there is no possibility of this as per the hypothesis of logistic regression expectation, given by (2).

$$0 \leq h_{\theta} \leq 1 \qquad (2)$$

Logistic regression is most commonly used with a categorical target variable and the data observed has a binary output, i.e. it belongs to either one class or another, i.e. 0 or 1. This algorithm basically gives the conditional probability that y belongs to a particular class given X input features. It has been observed to work well for smaller datasets.

#### 2. Random Forest

Random Forest is a bootstrapping algorithm that combines ensemble learning methods with decision tree framework to create multiple randomly drawn trees from the data, and averaging the results to output a result, which often leads to strong predictions.

#### 3. Decision Trees

Decision Trees are an approach used in supervised ML, a technique which uses labelled input and output datasets to train models. The approach is used mainly to solve classification problems, which is the use of a model to categorise or classify an object.

#### 4. Naive Bayes

This well-known classification technique is based on Bayes' theorem. It assumes that predictors are independent. It also assumes that whether or not one feature is present in a class is not at all related to whether another feature is present in it. Bayes' theorem helps to calculate the posterior (updated) probability as shown -

$P(c|x) = \frac{P(x|c).P(c)}{P(x)}$

Where,

- P(c|x) represents the updates probability of class (c, target) given predictor (x, attributes).
- P(x|c) represents the probability of the predictor given class.
- P(c) is the initial probability of class.
- P(x) is the initial probability of the predictor.

The Naïve Bayes model is easy to build and particularly useful for very large datasets [10]

### 5. K Nearest Neighbors Classifier

K-Nearest Neighbours (kNN) algorithm is also a supervised machine learning model used for categorisation problems. It does not make any postulations as to the the fundamental data distribution. It is known to perform well in pattern. recognition and predictive analysis. The classification process is as follows:

1) kNN first gathers data points that are close to the new data point taken into consideration. The distance between data points is effectively impacted by any attributes that can vary on a large scale.

2) Then, the algorithm sorts data points on the basis of their Euclidian distance from the new data point.

3) Next, the algorithm takes a specific number of data points, with lesser distance among all and then categorizes those data points. The number of closest data points is usually chosen as an odd number if the number of classes is 2. The class of the new data point will be the class with the highest number of data points.

### 6. Support Vector Machines

Support Vector Machine (SVM) is a coherent and simple yet highly preferred machine learning algorithm that can find its use in solving both classifications as well as regression problems. SVM is known for using little computational power to produce considerable accuracy. This algorithm aims to find such a hyperplane that distinctly classifies the data points in an N-dimensional space, where N is the number of features. A hyperplane is a decision boundary that classifies data points such that those on one side of the hyperplane belong to one 2 Authorized licensed use limited to: The University of Utah. Downloaded on December 19,2022 at 00:09:28 UTC from IEEE Xplore. Restrictions apply. class, while those on the other side belong to another class. A number of hyperplanes are possible to separate any two given classes. SVM finds such a plane that the distance between data points of both classes is maximum. This is called maximum margin and can be determined using the data points closest to the hyperplane. Such data points are called support vectors, and they influence the orientation and position of the hyperplane. The idea behind maximizing margin distance is that it adds to the expectation that test data points can be classified more accurately and confidently.

### 7. Bagged Decision Trees

Bagging on decision trees is done by creating bootstrap samples from the training data set and then built trees on bootstrap samples and then aggregating the output from all the trees and predicting the output.

### 8. AdaBoost

Adaptive boosting (most commonly seen in classification problems) combines multiple weak learners into one strong learner. First, equal weightage is assigned to all the data points and a decision stump is drawn out for a single input feature. Then, it analyses the obtained results and for any misclassified observations, it assigns higher weights to them. A new decision stump is then drawn which gives importance to the higher weights. Again, the miss-classified observations get the higher weights. This process keeps running in a loop till all observations are in the right place i.e all data points are classified correctly. AdaBoost is capable of being used in both regression and classification problems.

### 9. XGBoost

A more advanced version of Gradient boosting method called the XGBoost or Extreme gradient boosting comes under distributed machine learning. It focuses on model efficiency and computational speed. Where the older model's sequential analysis of datasets takes a long time, there is a need for this advanced algorithm to boost the performance of the model. To boost performance, it creates decision trees parallelly and implements distributed computing methods to evaluate complex models. The use of out of core computing helps analyse huge datasets and implements cache optimisation to utilize all resources and hardware.

### 10. Gradient Boosting

In gradient boosting, with every iteration, the overall model improves sequentially. In other words, base learners are sequentially generated such that the current base learner is always more effective than the previous base learner. The central idea in gradient boosting is to overcome the prediction of the previous learner by optimising its loss function, which is done by adding a new adaptive model that adds weak learners to reduce the loss function. There are three main components. First, the loss function which helps in reducing errors. Second, we need weak learners to compute predictions and to form strong learners. Third is the additional model that helps in regularising the loss function from the previous learner. Gradient boosting is an algorithm that can be used for both regression and classification problems.

### 11. Voting Classification

A voting classifier is a machine learning estimator that trains various base models or estimators and predicts on the basis of aggregating the findings of each base estimator. The aggregating criteria can be combined decision of voting for each estimator output.

### 12. Neural Networks

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

## B. Evaluation

The evaluation is based on Area Under ROC (AUC) curve, is the most commonly used measure in ML practice. It is a value between 0 and 1, and, the higher AUC, the better the predictive performance. It considers the cases of all possible thresholds that are used for (binary) classification, and calculates the area of the (TPR, FPR) curve of using these thresholds (TPR and FPR stands for True Positive Rate and False Positive Rate, respectively) as an overall measure of the model performance. Therefore, AUC is not restricted to the accuracy of any single threshold (e.g., 0.5 or 0). It is a comprehensive evaluation.

## IV. EXPERIMENTAL RESULTS

The results for the 10 algorithms implemented, with evaluation based on AUC curve, are outlined in Table 2.

Table 2

| Algorithm | Score |
|---|---|
| Logistic Regression | 0.88876 |
| Random Forest | 0.90725 |
| Decision Tree | 0.86560 |
| Naive Bayes | 0.66002 |
| K Nearest Neighbors | 0.66467 |
| Support Vector Machine | 0.59896 |
| Bagging Decision Trees | 0.88569 |
| AdaBoost | 0.92429 |
| XGBoost | 0.92028 |
| Gradient Boost | 0.92153 |
| Voting Classification | 0.86620 |
| Neural Networks | 0.84512 |

## V. CONCLUSION AND FUTURE SCOPE

In this project, we evaluated multiple machine learning algorithms. The algorithm that performed the best was AdaBoost with about 0.925 score on 50% of the test data. Due to this good performance, this model has the potential to be used in a variety of applications, including targeted marketing and credit risk assessment.

There are a few potential avenues for future work on this project:

- Further tuning of the model's hyperparameters may improve its performance.
- Adding additional features, such as information about an individual's investment portfolio or debts, may provide additional predictive power.
- Expanding the dataset to include a wider range of income levels and demographics could make the model more broadly applicable.

## ACKNOWLEDGEMENT

I would like to express my gratitude to my professor Shandian Zhe for introducing several fundamental and widely successful supervised/ unsupervised machine learning algorithms via CS5350/6350 Machine Learning, and to give me the opportunity to apply my knowledge on this Income Level Prediction project.