Proposal Project Report

Sai Chethan Singu
MSc. Data Analytics
(School Of Computing)
National College of Ireland
Dublin, Ireland
x18181937@student.ncirl.ie

I. Motivation

Income inequality is important key issue that many governments are trying to solve. Governments in different nations are using unique interventions or methods to address the profits inequality, a few of them are succeeding while the others are not. Reducing the differences in income will enhance and maintain the social developments between different communities. That will eventually improve the country's economic growth. This study focuses on the factors which are important in elaborating the income of the individuals. It primarily focuses on the important key aspects of income.

In this era, there are many issues can be solved with big data techniques. One such thing is that buying property. With machine learning methods, we can predict which property in which area is suitable to buy or sell. The term real estate is described as land and buildings, together with their natural resources such as crops, minerals or water. According to the predictions, the government and the developers can make important decisions about whether to develop the real estate in corresponding regions and also helps the customers in buying the right property at right time without loss and delay.

Road traffic accidents are the main cause of death globally as well as the eighth major cause of all deaths. The World health organization reported that every year 1.2 million deaths are happening due to road accident [1]. Factors influencing the road accidents are not only demographic and behavioral characteristics of an individual but also environmental factors and road conditions at the time of the incidence. This emphasizes the prediction of accidents and conditions that favours accidents for enhancing public transport.

II. Research questions

RQ.1 Determining the income level of an individual by using certain key attributes and to determining what are the important key attributes that have impact on the income.

RQ.2 Identifying the factors that influence customers to buy or sell properties based on factors like their location, land size, number of rooms and many others. Along with the prediction of the price also predicting the good price for which a property can be sold.

RQ.3 Predicting the severity especially in relation to climatic conditions for traffic accidents. And also analyzing important factors that have an impact on severity.

III. Initial Litrature Review

Researchers have made some attempts in the past to predict income levels using machine learning models. Prediction of the housing prices by using machine learning and data mining techniques in providing better insights to the income equality problem. They have used a gradient boosting classifier that was able to predict the highest accuracy when compared to other models [1]. Describes how to give a fast estimate of generalization error, but also to estimate variable importance and some important practical issues dealing with unequal sample sizes which predominantly helps in building the algorithm [2].

By seeing the historical property transactions in Australia, proposed a model by combining the stepwise and support vector machine based on mean squared error measurement which is a good approach for predicting buying and selling rates [3]. Predicted a model for raising and falling of house prices and analyze variance influence factors by two discrete values 0 and 1 as respective classes, where zero implies a decrease in house price and one implies an increase in house price [4].

Three tests were used for the model evaluation: out-of-bag error rate estimation, mean square error, and root mean square error. These are observed in the results which are obtained by using the random forest model [5]. Four specific methods of classification were evaluated for the severity of crashes. Investigated the effects of K-means clustering and latent class clustering methods on the efficiency of prediction of crash severity models [6]. The purpose of this work is to analyze with decision tree classifier, random forest and instance based learning methods with parameter k to predict and classify the severity of a motorcycle crash. The performances of these models were analyzed and compared to that of multinomial logit model [7].

IV. Data Sources

1. US Adults Income
   https://archive.ics.uci.edu/ml/datasets/Adult
   Description of dataset:
   Dataset consists of 30k rows which comprises 14 key attributes like age, work class, education, Education.num(1 to 16), Occupation, Relationship, Race, Sex, Income and many others.

2. Melbourne Housing Market
   https://www.kaggle.com/anthonypino/melbourne-housing-market
   Description of dataset:
   This Dataset gives us information about the factors that are influential in predicting the price of the house. There are 16 attributes which include Suburb, rooms, type, price, method, seller, date sold, land size, property count and many other features. It consists of approximately 10k rows.

3. US Countrywide Traffic Accidents
   https://www.kaggle.com/sobhanmoosavi/us-accidents
   Description of dataset:
   This dataset contains information about the various factors that should be taken into consideration to predict the severity of traffic which include temperature, wind, humidity, pressure and others. Dataset comprises of approximately 30k rows and 12 columns.

V. Identification of Machine Learning Methods

Adult census Income

Decision Tree: It is one of the supervised learning algorithms. The aim of using a Decision Tree is to build a training model that can be used to predict the target variable value by learning basic rules of judgment inferred from the training data. Each internal node represents attribute and leaf node denotes class label. The reason behind choosing this method is that by handling both continuous and categorical variables, it is easy to predict the most influential factors like income and other factors.

Logistic Regression: Logistic regression is a classification algorithm model. Based on probability, analysis is done in this type of regression. We use the sigmoid function which ranges between 0 and 1 to map the predictions with respect to probabilities. The rationale behind the algorithm is either to predict the income is >50k or <=50k which falls under binary output can be predicted by logistic regression.

Melbourne Housing Market

Linear Regression: Linear Regression is a supervised machine learning algorithm where the predicted output is continuous. It is used to predict values within a continuous range rather than trying to classify them into categories. In this method, the dependent variables are always continuous, the independent variables can be continuous or discrete. The formula is: y=mx+c. The rationale for choosing linear regression is that it can handle feature selection which helps for this dataset.

US Countrywide Traffic Accidents

Random Forest: It is an ensemble learning method based on classification and regression trees. Random forest algorithm first creates the decision trees based on sample data and then predicts from each of those trees and finally choose the best possible solution with the help of voting system. And also reduces the overfitting problem by taking the average. The rationale for choosing this algorithm is that the target variable is binary and it also eliminates the overfitting problem. And attributes are also uncorrelated to each other.

KNN (K- Nearest Neighbors): KNN algorithm is used for the classification as well as regression models. KNN algorithm assumes that similar things are near to each other. In other words, it is based on feature similarity. It classifies based on distance functions such as Euclidean distance, Manhattan distance, Minkowski distance and others. In this study, Euclidean distance function is used. The rationale behind this algorithm is that there is no linearity between the attributes and we can weigh the categories differently and predict the severity of the traffic accurately.

VI. Identification of Evaluation Methods

Accuracy:
Accuracy is one of the important factors to be evaluated when using a classification algorithm. It is the ratio of the number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Precision:
Precision is determined by dividing true positive values with the sum of true positive and false positive.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Sensitivity or Recall:
Calculated by dividing the true positive with the sum of true positives and the false negatives.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

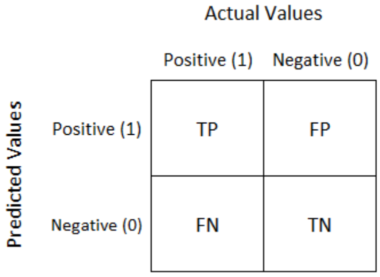
F1 Score:

F1 score is the harmonic mean of precision and recall. Precision tells about actual positive out of the total predicted positive. Recall tells about actual positive from the total actual positive.

F1 Score: 2(Precision * Recall)/ (Precision + Recall)*

Confusion Matrix:

The performance of classification model is explained by the confusion matrix, where the output can be of two or more categories. It helps to visualize the algorithm performance.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Mean Squared Error:

It calculates the average square of the errors between the actual and the predicted values.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error:

It is the distance between predicted and observed values and it is the square root of the mean square error.

Correlation:

A correlation could be positive, denotes that both variables are related, or negative, meaning that when the value of one variable increases, the values of the other variables fall.

Correlation can also be zero, which means that the variables are unrelated to each other.

Euclidean Distance:

The formula for measuring the distance between any two data points in a plane.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

VII. BIBLIOGRAPHY

[1] World Health Organization. https://www.who.int/

[2] Chakrabarty and Biswas: A Statistical Approach to Adult Census Income Level Prediction.

[3] Richard , Stevens: An extension of Breiman’s bagging idea and developed as a competitor to boosting model.

[4] Danh Phan, “Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia”, 2018. International Conference on Machine Learning and Data Engineering (ICMLDE).

[5] Banerjee, Dutta: Predicting Housing Price Direction using Machine Learning. 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI).

[6] Lee, Yoon 2, Kwon, Jongtae: Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study, Applied Sciences Published: 23 December 2019.

[7] Iranitalab, Khattak: Comparison of four statistical and machine learning methods for crash severity prediction. Published by Elsevier.

[8] Wahab, Jiang: A study on machine learning based algorithms for prediction of motorcycle crash severity.