7CS030 – Concepts of Artificial Intelligence & Technologies

School of Mathematics and Computer Science Faculty
of Science and Engineering

Coursework Individual Assignment

Kinza Waheed
2313970
17-11-2023

Table of Contents

Abstract

This report contains the analysis of three data sets; prediction of house prices, prediction of a player who will last 5 years in the NBA and country data set. Multiple Machine Learning models have been implemented to gain successful results. Regression model has been used to predict the house price. Clustering has been performed on the country data set to observe the relationship between variables and neural network model has been implemented on NBA rockie performance data set.

Keywords: Regression, Clustering, Supervised and Unsupervised Learning, Neural Networks, Ethics of AI

Task 1

The house price data set contains information about the house sales in King County, USA. It has 18 columns such as price, number of bedrooms, bathrooms, floor etc. Regression Model has been implemented on this data set to predict the price of the house. Regression is a widely used supervised learning approach in the machine learning sector. Regression Models are usually used for forecasting and prediction. Its purpose is to construct a model that can accurately estimate the target variable (price) based on the independent variables.

The most common and simplest type of Regression model is Linear Regression model which assumes a linear relation between target variable and independent variables also knows as predictor variables. To run this analysis, various packages have been used in python such as pandas, numpy, matplotlib and scikit-learn. A well-known open-source Python machine learning library is called Scikit-learn.

Linear Regression Model

Linear Regression has been implemented on the dataset with the target variable house prices and independent variable squarefoot living. Firstly, one fourth dataset has been trained by the code **X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/4, random_state = 0).** For this, *train_test_split* function has been imported from *sklearn.model_selection*. Secondly, linear regression model fits into the train data set by the code *regr = LinearRegression() and regr.fit(X_train, y_train)*. Thirdly, to obtain the accuracy of the model coefficients, intercept, mean squared error and coefficient of determination has been obtained. To visualize the results, a scatter plot has been made using the *matplotlib.pyplot* lib library in python.

Figure 1

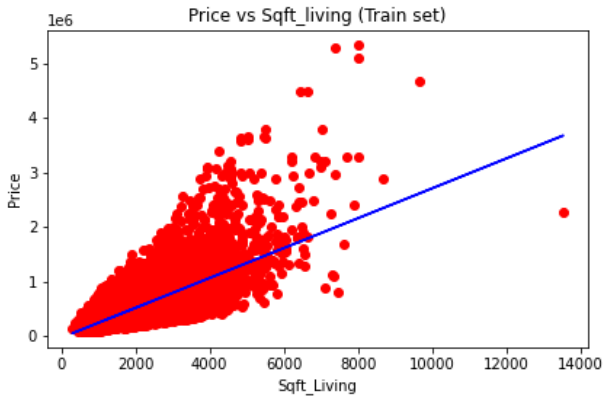

Figure 2

In figure 1, the scatter plot shows that there is a positive strong relationship between the price of house and square foot living, which means that houses with more living room space have more prices. However, the correlation is not perfect, which suggests that there are other factors which influence the price of the house. In addition, Figure 2 shows the scatter plot of train data set with the regression fitted line.

The fitness of the linear regression model has been assessed through the following points.

- The square foot living variable has a coefficient of about 273.98. This shows that, on average, the estimated cost of the home rises by about $273.98 for every square foot of extra living area.
- R-squared, or coefficient of determination, is equal to 0.50. This indicates that the linear relationship between square foot living and house price variability accounts for almost half of the variation in house prices.

This model has some **limitations**, as from the results, we can conclude that 50% of the variability in house prices can be explained by the linear relationship with square foot living. Multiple Regression model would be a good fit that could account for the other factors as well to predict accurate house prices.

Multiple Linear Regression Model

This model works the same as linear regression model, except that it can include more than one independent variable. Here we have used all the variables from the data set except sqft_living15 and sqft_loft. Feature selection process has been done using heat map plot.

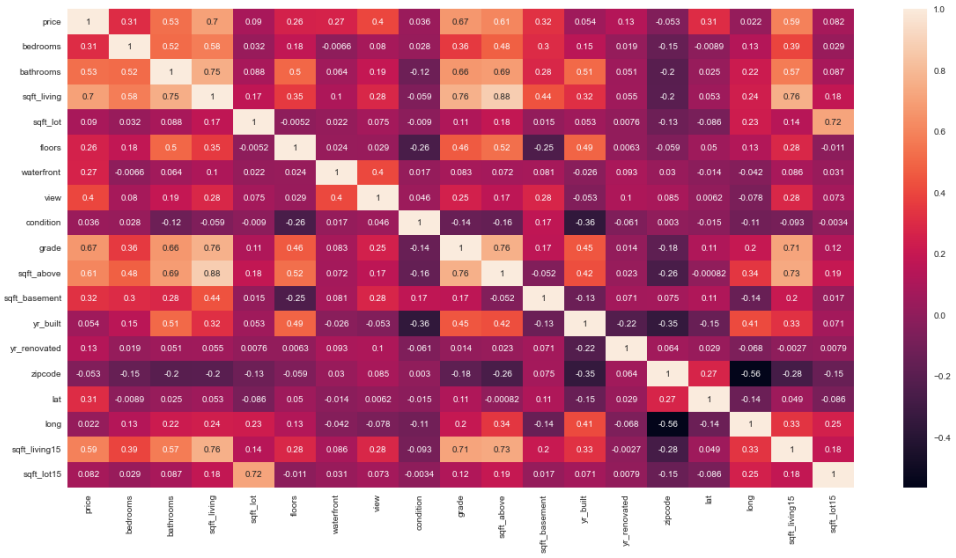

Figure 3

Figure 3, suggested that price and square foot living has 0.70 correlation. Moreover, there is a strong correlation between sqft_ living, sqft_living15 and sqft_loft. So, by removing highly correlated variables, the accuracy of the model improved from 0.66 to 0.70. The results showed that this model can predict 70% of the house prices based on its features like bedroom, bathroom, square foot living, grade etc.

Task 2

Unsupervised Learning Algorithm

Unsupervised algorithm is a branch of machine learning where models are not supervised using training data set. However, the model itself requires finding the hidden patterns and insights in the data set. Unlike supervised approach, where models are trained using labeled data under the supervision of training data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format (Anon., n.d.)

Clustering is widely used unsupervised learning technique, where the model automatically makes clusters of the data set based on their similar characteristics. It focuses on the features of all input data sets rather than focusing on output data like regression.

Country Data Set

In the country data set, we have 167 countries and their health, imports, exports, gdpp, child mortality, inflation, income, life expectancy and total fertility rate data. In this report, clusters have been made to analyse the relationship between some features and group some data points together based on their similarities. *Kmeans* function imported from *sklearn.clusters* library to implement the K means clustering algorithm on the country data set. It is a centroid based algorithm, where each cluster is associated with the centroid. The data points in a cluster have a minimum distance with the centroid of that cluster compared to the distance with another cluster centroid. The distance of data points with the centroid is calculated by a mathematical formula called *Euclidean distance function*.

Figure 4 represents the pairplot of the country data set that shows the pairwise relationship among the variables and histogram for each variable. It helps with the feature selection of clustering and made interesting findings such as

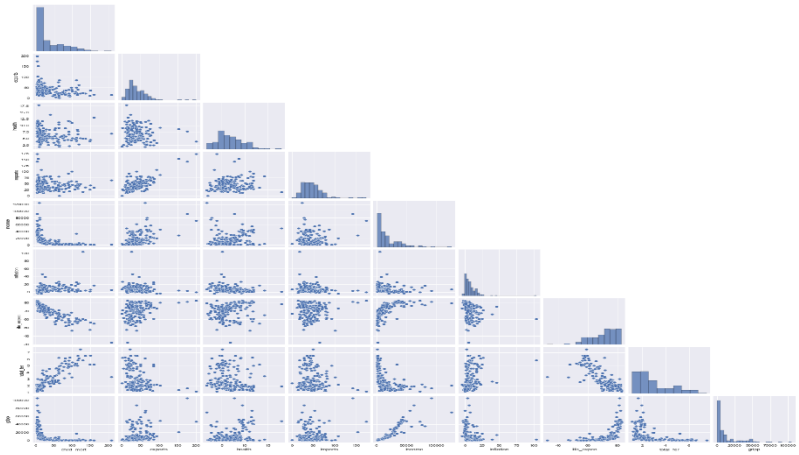

Figure 4

- Countries with high GDP per capita have lower fertility rate also child mortality and higher life-expectancy. As child mortality and fertility rate are directly proportional.
- Countries who spend more on heath, have higher life expectancy
- GDP per capita increases as imports, exports and income of the country increases.
- Countries have low fertility rate when they spend more on health.

Hence, we can conclude that there is a strong relationship between economic development and social wellbeing.

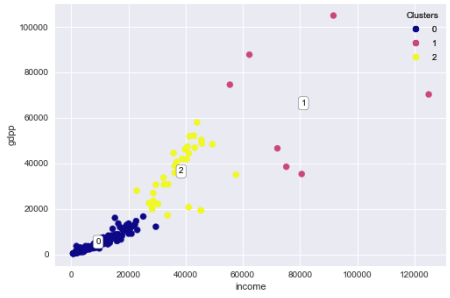

Figure 5

In the figure 5, gdpp and income feature has been selected from the country data set. Four clusters made by the code *model = KMeans(n_clusters = 3, n_init='auto', random_state=5). Countries with similar features like high gdpp or high income have grouped together as cluster one(purple) which represents that they are developed countries. Moreover, based on the lower gdp per capita and low income, this model made a cluster in blue representing under-devel****oped countries.*** There is a capacity to make more clusters in between blue and yellow as we can see the overlapping of data points.

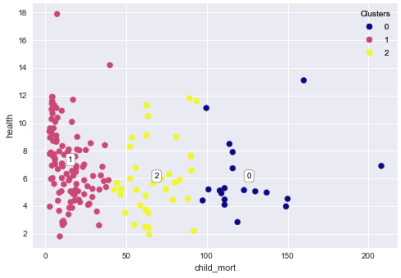

Figure 6

Another, insights have been discovered from K means clustering algorithm. Figure 6, shows that there are some countries whose child mortality rate is high as there spending on health is low. However, countries with the high health expenditure tend to have less child_motality rate.

Analysis of GDP, Child Mortality and Income with the Countries Development Status Using Boxplot

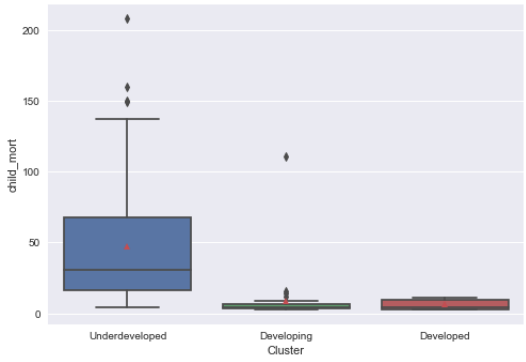

Figure 7

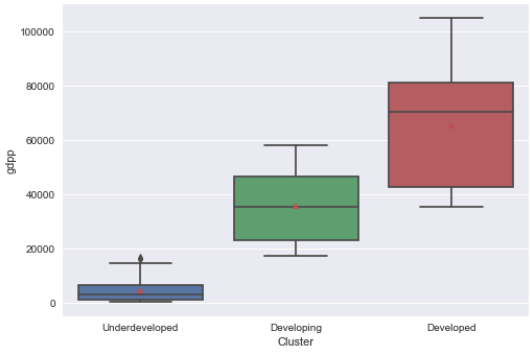

Figure 8

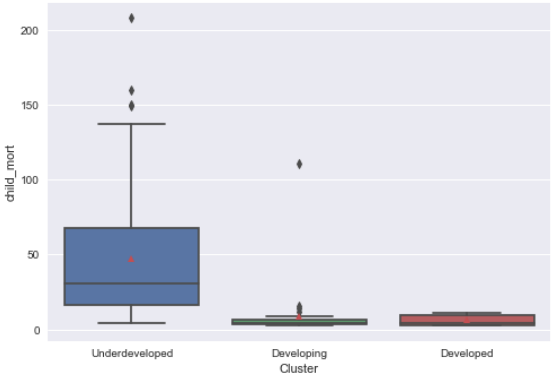

Figure 9

*F*igure 7,8 & 9 shows that box plot of child mortality,gdp per capital and income per person in respect to the countries development status.

The following observations has been made from the box plot.

- Developing countries have Medium GDPP, medium Income and mild child mortality rate.
- Developed countries have High GDPP, High income and very low child mortality rate.
- Under-Developed countries have Low GDPP, Low income and very high mortality rate and should be our primary focus.

Task 3

In this data set of players, we have done analysis to predict whether a player will last 5 years in NBA. The dataset contains 18 features with the target variable **Target_5Yrs** with 1: if career length >= 5 yrs or 0: if career length < 5 yrs. We have run Logistic Regression, Gaussian NB and Neural Network Model to make this prediction. The results of the analysis are below.

Logistic Regression Model

Logistic Regression Model is a type of statistical model which is often used for binary classification and prediction. It estimates the probability of an event occurring, in this case logistic regression algorithm has been used to predict the probability of an NBA player lasting 5 years in the league. It is a supervised machine learning approach which means it learns from the labelled data set. Logistic regression derived from the logistic function, which is also called sigmoid function, that squishes the linear prediction (that has been fitted in the data points as linear model), into a range between 0 and 1. In this dataset, the model calculates the probability score of each player based on their features such as ….,….. and then classified players as they last 5 years as 1 or if not then 0.

The accuracy of the model is 0.69, it means that this model predicts the player's career longevity very well. Based on the data points of the predictor variable, we can now predict the career longevity of a new player.

Gaussian Naïve Base

GaunnianNB is another supervised machine learning algorithm, implements the Gaussian Naive Bayes algorithm for classification. This model works on the Naïve base theorem with an

assumption of independence among predictor variables or conditional independence between every pair of variables given the value of the class variable In this case Target_5Yrs.

To run the Gaussian Naïve Base on the dataset, GaussianNB has been imported from sklearn.naive_bayes. Data has been normalized to gain the best results out of this model. Train test the data first and then gain the accuracy of the model by this code print('Our accuracy is %.2f:' % gnb.score(X_test, y_test)), which showed 0.63. It means that this model is not bad, but from the logistic regression we gained 69% accuracy, so the model make the classfication of data into career length >= 5 yrs or career length < 5 yrs more perfectly.

Neural Networks

Neural networks also known as artificial neural network ANN are a subset of machine learning called deep learning. This model is inspired by the human brain and mimicking the communication between biological neurons (IBM, n.d.)It compromises of three layers: first, input layer where information from the data enters into arificial neural network, it further analyse or preprocess the data. Second, comes the hidden layers which take input from the input layer or other hidden layers. Deep neural networks could have several hidden layers with millions of nodes linked together (Aws, n.d.)Each neuron or nodes contains an activation function which calculates the input data, assigned weight to the result and pass it to another hidden layer node. Finally, it gives output in the last layer.

Here to analyze the rookie NBA data to predict the probability of an NBA player lasting 5 years in the league, we have used ***relu*** activation function in the nodes and four hidden layers with ***hidden_layer_sizes=(20,60,30,60)***.

Neural network model implemented in this data set (selected all features of data set) by importing MLPClassifier from sklearn.neural_network. In MLP classifier we could use 4 activation formulas which are relu, tanh, identity and logistic. After using all these activation formulas, relu works best for this data set, giving accuracy of 0.68. This showed that predictor variables are defining 68% if the player can last upto 5 years in the league.

Heat Map for feature selection process

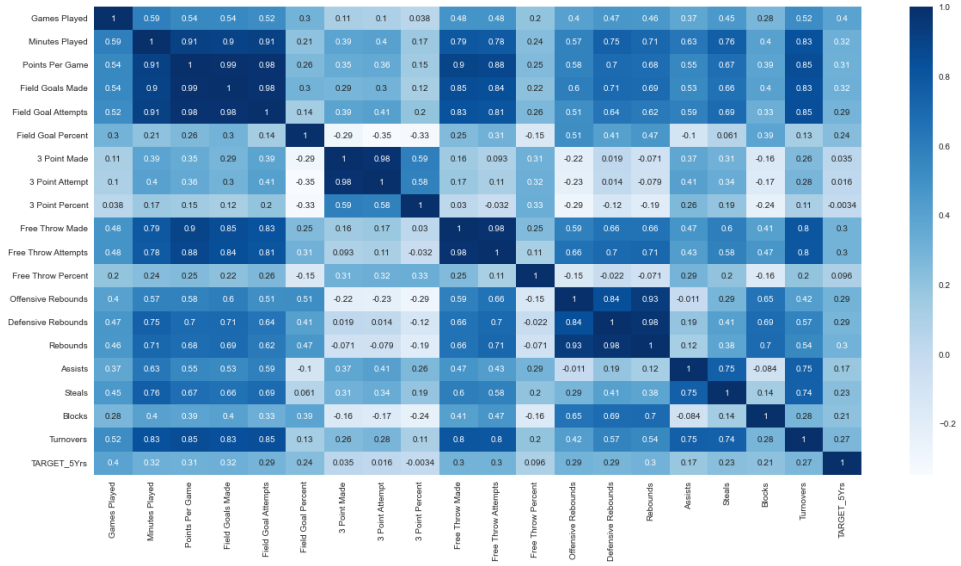

Figure 10

Improvement of Model

Moreover, by using the heatmap, I have dropped some features in the data set like 'Field Goals Attempts', 'Field Goal Made', '3Point Attempt', 'Free Throw Made ' and 'Defensice Rebounds and run neural network model, logistic regression model and gaussian naïve base, rest is all same. The results give me 100% accuracy. It means all three models are the best fit model to predict the NBA player career longevity with the perfect feature selection.

Model	Accuracy (with all features)	Accuracy (after Feature Selection)
Logistic Regression	0.69	1.00
Gaussian Naïve Biase	0.64	1.00
Neural Networks.	0.68	1.00

Task 4

Trolley Problem (Ethics in AI)

A new era of transportation has begun with the incorporation of artificial intelligence (AI) in autonomous vehicles, posing ethical challenges to society that call for careful deliberation. Of these difficulties, the Trolley Problem is particularly noteworthy as a paradigmatic ethical conundrum that lies at the interface of technology and morality and goes beyond conventional philosophical discourse. The Trolley Problem is never about saying what the right answer is, it's about how we reason morally (Greene, 2021).

Researchers and policymakers rely on an abundance of interdisciplinary information to navigate this ethical situation. The design and programming of autonomous cars are influenced by opposing moral frameworks such as utilitarianism and deontology.

Utilitarian Theory

The utilitarian theory of ethics, which holds that the ideal course of conduct is the one that maximizes general happiness or well-being (Edward N. Zalta, 2014), is frequently examined through the lens of the trolley problem. A utilitarian would argue that, in the trolley scenario, pulling the lever is morally acceptable because it would only cause the death of one person rather than five.

Deontologist Theory

Even though pulling the lever would save more lives, some ethical theories contend that it is unethical to do so. Deontologists, for instance, hold that some behaviors are always morally justified or immoral, regardless of the outcomes. A deontologist would contend that in the trolley situation, killing one worker is wrong even if it saves five others.

AI ethics is a challenging rapidly developing field. In conclusion, the complex relationship between ethics and technology is best illustrated by the Trolley Problem in the context of autonomous vehicles. It takes a deep grasp of psychological quirks, society norms, and philosophical ideas to successfully navigate this ethical terrain. The question of whether pulling the lever is morally acceptable has no simple solution. However, the dilemma can be used to investigate a few significant ethical questions, including deontology, the nature of utilitarianism, and the worth of human life.

Bibliography

Anon., n.d. [Online]
Available at: https://www.javatpoint.com/unsupervised-machine-learning

Anon., n.d. *IBM- What are Neural Networks?.* [Online]
Available at: https://www.ibm.com/topics/neural-networks

Aws, n.d. *What is a neural network.* [Online]
Available at: What is a Neural Network? - Artificial Neural Network Explained - AWS (amazon.com).

Edward N. Zalta, U. N., 2014. The History of Utilitarianism. *The Stanford Encyclopedia of Philosophy.*

Greene, J., 2021. 176: Joshua Greene on Morality, Psychology, and Trolley Problems. June.

IBM, n.d. *What are neural networks.* [Online]
Available at: https://www.ibm.com/topics/neural-networks