University of Essex

University of Essex

Department of Mathematical Sciences

CE802 : MACHINE LEARNING

Title of Your Project:Machine Learning Project

Your Name: Alankrita Mitra

Supervisor: **Your Supervisor:Dr.Vito De Feo**

January 12, 2023
Colchester

Machine Learning Project

Abstract

Machine learning algorithms has been performed in this particular task that is provided. Therefore, the energy business would want to determine, based on a few key criteria, if the client is having trouble paying the cost of power and we have performed the classification task here. We have done regression task in other part where In case of regression problem The corporation wishes to create a different approach to forecast the change in a customer's annual expenses as a result of an increase in energy costs.

We have used machine learning algorithms like linear regression, decision tree, support vector machine and logistic regression in our problem. R squared is used to see which regression algorithm performs the best and accuracy is used as parameter in determining which algorithm performs best in classification task.

Linear regression model with R squared value of 0.693 performs the best where as decision tree has the accuracy of 0.838 performs the best in classification problem.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

A subfield of computer science and artificial intelligence known as "machine learning" focuses on using data and algorithms to simulate human behaviour and steadily increase accuracy. With the used data, we may carry out tasks like categorization and prediction. Learning algorithms are defined as those that raise performance metrics as we do a job. We can divide the machine learning algorithm in to supervised, unsupervised and reinforcement learning.

- Supervised learning algorithm: When using supervised learning algorithms, which primarily rely on labelled input and output of the training data, these algorithms are more accurate than unsupervised learning algorithms since they require human interaction. In supervised learning, a training set is used to teach models to produce the desired output. In this training dataset, the inputs and outputs are precise, which enables the model to develop over time.

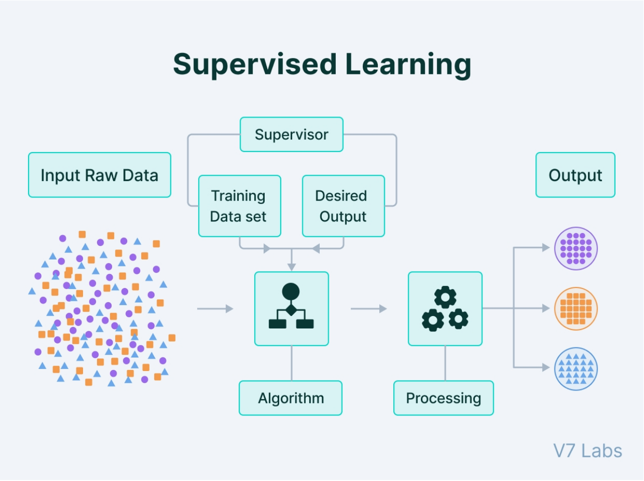

Figure 1.1: Supervised learning algorithm(https://towardsdatascience.com)

- Unsupervised learning algorithm: Unsupervised learning algorithms rely on unlabeled or raw data to carry out their work, hence they are less accurate than supervised learning algorithms since they don't need human interaction. In mathematics, unsupervised learning is the process of seeing several instances of a vector X and learning the probability distribution p(X) for these instances.

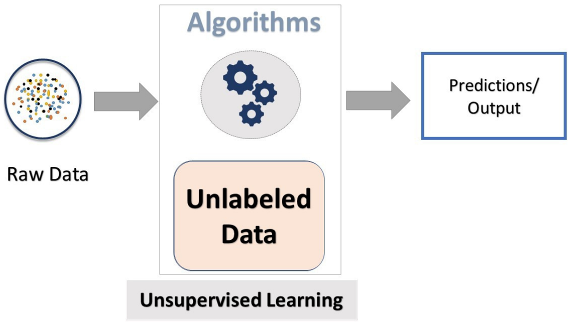

Figure 1.2: Supervised learning algorithm(https://towardsdatascience.com)

- Reinforcement learning: Similar to unsupervised learning, reinforcement learning does not need tagged data. Alternatively, a model develops through time through interaction with its surroundings.

CHAPTER 2

Data Preprocession

Before building any model in order to increase the performance of it we need to do all the data pre-processing task.We have given 2 dataset one is P2 and other P3 which contains 1000 rows 22 columns and 1500 rows and 37 columns respectively.

- The dataset used contains missing value so we fill the missing values with mode technique.
- The dataset is divided into training and testing part. 70 percent of the data is divided into training and remaining 30 percent into testing part.
- Normalization of data is done using MinMax Scalar funtion. Feature Scaling using Normalisation (i.e. Min Max Scaler),as regression based algorithms which use gradient descent for optimisation require scaled data.
- Using heatmap we have checked whether there is any presence of multi-collinearity or not.
- Using dummy variable we have converted the categorical columns into numerical one.

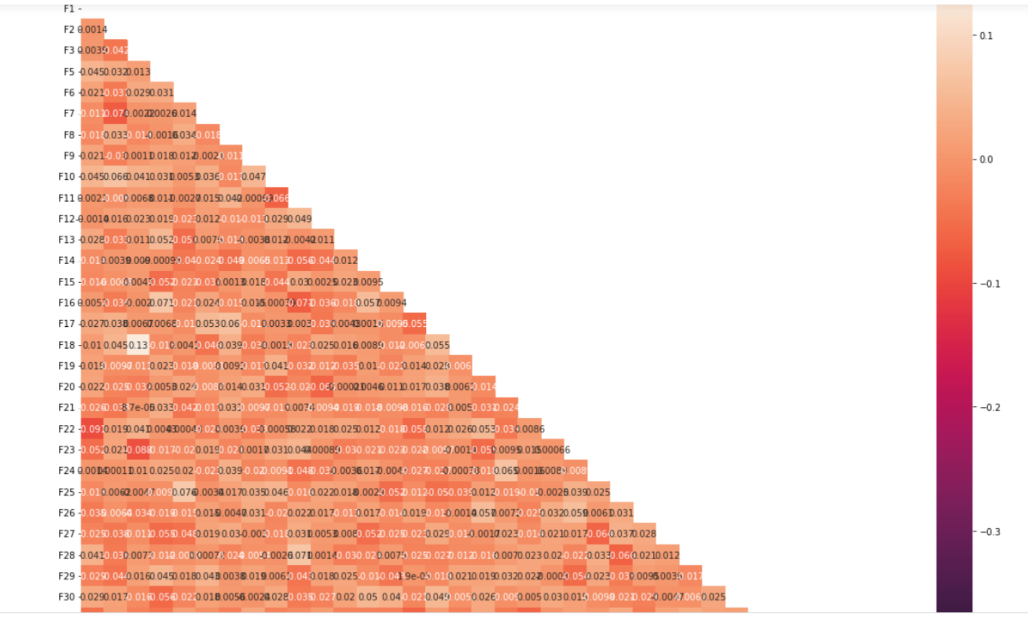

Figure 2.1: Heatmap(source code)

Heatmap is used to see the existence of multicollinearity which can destabilize the model.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc26420cfa0>
```

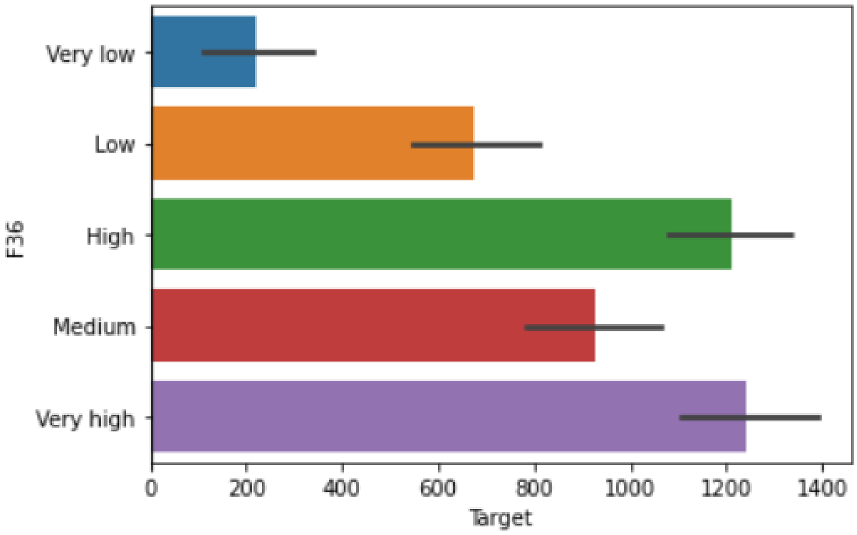

Figure 2.2: Pairplot(source code)

Bar graph is plotted between a categorical variable and a target variable.

CHAPTER 3

Methodology

Machine learning algorithm is used in order to find which model performs the best in both the regression and classification task.

Logistics Regression: Analytics classification issues are a significant subset of issues where the response or outcome variable has discrete values. Logistic regression is a method for modelling the probability of a discrete output given an input variable. Logistic regression determines the chance that a specific event, like voting or not voting, will take place based on a set of independent factors and a dataset. The outcome is a probability since the dependent variable's range is 0 to [McCallum, 2019][Tolles and Meurer, 2016].

Table 3.1: Hyperparameter tuning used (Logistic Regression)

Hyperparameter used	Values
Solver	['newton-cg', 'lbfgs','liblinear']
Penalty	['l2','l1']
'C'	[100, 10]
random state	[42]

Support Vector Machine:It is a set of supervised learning algorithm used for regression, classification and detecting outliers.[McCallum, 2019]. Support vector machine is used because it has following advantage:

- In high dinmensional spaces support vector machine works perfectly also.
- Even when there are more dimensions than samples, the method is still successful.
- It is also memory efficient since it only uses a portion of the training data for the decision function (known as support vectors).

Table 3.2: Hyperparameter tuning used (Support vector machine)

Hyperparameter used	Values
kernel	['poly', 'rbf']
gamma	['scale']
'C'	[50, 10, 1.0]
random state	[42]

Decision Tree classifier: Models for regression or classification are built using Decision Tree in the form of a tree structure. As a dataset is divided up into ever smaller subgroups, a decision tree is gradually developed to go with it. It is a group of divide-and-conquer problem-solving strategies that makes use of a tree-like structure to predict the value of an output variable.

Table 3.3: Hyperparameter tuning used (Random Forest Method)

Hyperparameter used	Values
max_depth	np.arange(2,20)
min_samples_split	np.arange(2,10,2)
max_samples_split	np.arange(1,11,2)

Linear Regression: The goal of linear regression is to minimise the residual sum of squares between the targets that were seen in the dataset and the targets that the linear approximation anticipated. By analysing the data and fitting a linear equation to it, linear regression makes an attempt to describe the connection between two variables. A dependent variable is one that is used to explain another variable, while an explanatory variable is one that does the opposite. If there is a link between the variables of interest, it should be established before attempting to fit a linear model to the observed data.

CHAPTER 4

Result

Table 4.1: classifcation report of Decision Tree Classifier

Classification Report	Precision	Recall	F1-Score	Support
False	0.83	0.78	0.81	193
True	0.81	0.85	0.83	207
Accuracy			0.82	400
macro avg	0.82	0.82	0.82	400
weighted avg	0.82	0.82	0.82	400

Table 4.2: classifcation report of Support vector machine

Classification Report	Precision	Recall	F1-Score	Support
False	0.60	0.82	0.69	193
True	0.75	0.48	0.59	207
Accuracy			0.65	400
macro avg	0.67	0.65	0.64	400
weighted avg	0.67	0.65	0.64	400

Table 4.3: classifcation report of Logistic Regression Method

Classification Report	Precision	Recall	F1-Score	Support
False	0.71	0.69	0.70	193
True	0.71	0.73	0.73	207
Accuracy			0.71	400
macro avg	0.71	0.71	0.71	400
weighted avg	0.71	0.71	0.71	400

Table 4.4: Cross validation accuracy comparison

Algorithm	CV best score
Decision Tree	84.33(percent)
Support vector machine	66.33(percent)
Logistic Regression	72.16(percent)

Table 4.5: Regression result

Algorithm	R squared	MAE	MSE	RMSE
Decision Tree(without hyperparameter tuning)	0.175	801.05	1215047.069	1102.29
Decision Tree(with hyperparameter tuning)	0.4280	690.41	842819.20	918.05
Linear Regression	0.693	542.27	451367.89	671.83
Random forest method	0.663	522.32	495942.11	704.23

CHAPTER 5

Discussion

We have compared the result and found out which algorithm performs the best. There are 3 table 4.1, 4.2, 4.3 there we have used accuracy as a parameter in order to compare the best parameter since it is a classification problem. Decision tree performs the best with accuracy of 0.82 where as other algorithms like support vector machine and logistic regression have 0.64 and 0.71 percent accuracy

We have found out the co-efficient of determinant value in the regression problem which is taken as base parameter for comparing which model performs the best. Linear regression has the best r squared value of 0.693.

CHAPTER 6

Conclusion

Machine learning algorithm is suitable for above task and we ca perform both regression and classification problem also. Linear regression perform the best in case of regression problem and decision tree performs the best in case of classification problem.

Bibliography

[McCallum, 2019] McCallum, A. (2019). Graphical models, lecture2: Bayesian network represention. *PDF). Retrieved*, 22.

[Tolles and Meurer, 2016] Tolles, J. and Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama*, 316(5):533–534.

- Pandian, B. Jaganatha; Noel, Mathew Mithra (2018-09-01). Control of a bioreactor using a new partially supervised reinforcement learning algorithm. Journal of Process Control. 69: 16â29. doi:10.1016/j.jprocont.2018.07.013. ISSN 0959-1524. S2CID 126074778

- van Otterlo, M.; Wiering, M. (2012). Reinforcement learning and markov decision processes. Reinforcement Learning. Adaptation, Learning, and Optimization. Vol. 12. pp. 3â42. doi:10.1007/978-3-642-27645-3-1. ISBN 978-3-642-27644-6.

- https://www.wikipedia.com

- https://www.scikit-learn.org

- A Supervised Learning Model for Detecting Alzheimer′s Disease Paperback â 7 Nov. 2017 by C. R. Aditya (Author), M. B. Sanjay Pande (Author)

- Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) Foundations of Machine Learning, The MIT Press ISBN 9780262018258.

- Stuart J. Russell, Peter Norvig (2010) Artificial Intelligence: A Modern Approach, Third Edition, Prentice Hall ISBN 9780136042594.

- Machine Learning using python by U Dinesh Kumar.

- https://www.collectcourses.com