**MARKS PREDICTION**

Mohd Amaan (11910898)

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab, India

[mohdamaan5668@gmail.com](mailto:mohdamaan5668@gmail.com)

**Abstract-While performing tests on my topic which is Marks Prediction, I did experiment with a real world dataset, and to explore how machine learning algorithms can be used to ease our work and to predict the new output with the help of the given data. I was expected to gain experience using a common machine learning library and was expected to submit a report and a research paper about the dataset and the algorithms used. After performing the required tasks on a dataset of my choice, herein lies my research paper.**

**I Introduction**

Machine learning is a sub-domain of computer science which evolved from the study of pattern recognition in data, and also from the computational learning theory in artificial intelligence. It is the first-class ticket to most interesting careers in data analytics today. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions

Machine Learning can be thought of as the study of a list of sub-problems, viz: decision making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic a algorithms, sparse dictionary learning, etc .Supervised learning ,or classification is the machine learning task of inferring a function from a labelled data In Supervised learning,we have m training set, and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a functioa that maps the input vector to the output vector with minimal error. In an optimal scenario, a model trained on a set of examples will classify an unseen example in a correct fashion, whica requires the model to generalize from the training set in a reasonable way. In layman's ternus, mpervised learning can be termed as the process of concept learning, where a brain is exposen to a set of inputs and result vectors and the brain learns the concept that relates said inputs to output A wide array of supervised machine learning algorithms are available to the machine learning enthusiast, for example Neural Networks Decision Trees, Support Vector Machines Random Forest, Naive Bayes Classifier, Bayes Net, Majerity Classifier(4.7.8.9) etc. and they each have their own merits and demerits. There is no single algorithm that works for all cases, as merited by the No free lunch theorem. In this project, I tried and found patterns in a dataset. which is a sample of marks obtained by studying for certain number of hours, and attempt to throw various intelligently-picked algorithms at the data, and see what sticks.

**Classification** is about predicting a label, by identifying which category an object belongs to based on different parameters.

**Regression** is about predicting a continuous output, by finding the correlations between dependent and independent variables.

**Chart, scatter chart

Description automatically generated**

**II Linear Regression**

Linear Regression is known as one of the simplest Machine learning algorithms that branch from Supervised Learning and is primarily used to solve regression problems.

The use of Linear Regression is to make predictions on continuous dependent variables with the assistance and knowledge from independent variables. The overall goal of Linear Regression is to find the line of best fit, which can accurately predict the output for continuous dependent variables. Examples of continuous values are house prices, age, and salary.

Simple Linear Regression is a regression model that estimates the relationship between one single independent variable and one dependent variable using a straight line. If there are more than two independent variables, we then call this Multiple Linear Regression.

Using the strategy of the line of best fits helps us to understand the relationship between the dependent and independent variable; which should be of linear nature.

**The Formula for Linear Regression**

If you remember high school Mathematics, you will remember the formula: ***y = mx + b*** and represents the slope-intercept of a straight line. ‘y’ and ‘x’ represent variables, ‘m’ describes the slope of the line and ‘b’ describe the y-intercept, where the line crosses the y-axis.

For Linear Regression, ‘y’ represents the dependent variable, ‘x’ represents the independent variable, 𝜷0 represents the y-intercept and 𝜷1 represents the slope, which describes the relationship between the independent variable and the dependent variable.

Diagram, schematic

Description automatically generated

**III Literature Review**

A core objective of a leamer is to generalize from its experience. The computational analysis of machine leaming algorithms and their performance is a branch of theoretical computer science known as computational learning theory. Because training sets are finite and the future uncertain, learning theory usually does not yield guarantees of the performance of algorithms. Instead, probabilistic bounds on the performance are quite common. The bias variance decomposition is one way to quantify generalization error.

For the best performance in the context of generalization, the complexity of the hypothesis should match the complexity of the function underlying the data. If the hypothesis is less complex than the function, then the model has underfit the data. If the complexity of the model is increased in response, then the training error decreases. But if the hypothesis is too complex.

the the model is subject to overfitting and generalization will be poorer

In addition to performance bounds, learning theorists study the time complexity and feasibility of learning In computational learning theory, a computation is considered feasible if it can be done in polynomial time. There are two kinds of time complexity results. Positive results a show m that a certain a class n of functions can be learned in polynomial time. Negative results show that certain classes cannot be learned in polynomial time.

**IV** **Technology Implemented**

**Python**

Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for an emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

**Features**

1.Interpreted

In Python there is no separate compilation and execution steps like C/C++. It directly run the program from the source code. Internally, Python converts the source code into an intermediate form called bytecodes which is then translated into native language of specific computer to run it.

2.Platform Independent

Python programs can be developed and executed on the multiple operating system platform. Python can be used on Linux, Windows, Macintosh, Solaris and many more.

3.Multi-Paradigm

Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect oriented programming.

4.Simple

Python is a very simple language. It is a very easy to learn as it is closer to English language. In python more emphasis is on the solution to the problem rather than the syntax.

5.Rich Library Support

Python standard library is very vast. It can help to do various things involving regular expressions. documentation generation, unit testing, threading, databases, web browsers, CGI, email, XML, HTML, WAV files, cryptography, GUI and many more.

Free and Open Source

Firstly, Python is freely available. Secondly, it is open-source. This means that its source code is available to the public. We can download it, change it, use it, and distribute it. This is called FLOSS (Free/Libre and Open Source Software). As the Python community, we're all headed toward one goal - an ever-bettering Python.

**V** **Data Preprocessing, Analysis & Visualization**

Machine Learning algorithms don't work so well with processing raw data. Before we can feed such data) to an ML algorithm, we must preprocess it. We must apply some transformations on it. With data preprocessing, we convert raw data into a clean data set. To perform data this, there are 7 techniques

1.Rescaling Data

For data with attributes of varying scales, we can rescale attributes to possessi same scale. We rescale attributes into the range 0 to 1 and call it normalization. We use the MinMaxScaler class from scikit learn. This gives us values between 0 and 1.

2. Standardizing Data

With standardizing, we can take attributes with a Gaussian distribution and different means and standard deviations and transform them into a standard Gaussian distribution with a mean of 0 and a standard deviation of 1.

3. Normalizing Data

In this task, we rescale each observation to a length of 1 (a unit norm). For this, we use the Normalizer class.

4. Binarizing Data

Using a binary threshold, it is possible to transform our data by marking the values above it I and those equal to or below it, 0. For this purpose, we use the Binarizer class

5. Mean Removal

We can remove the mean from each feature to center it on zero

6. One Hot Encoding

When dealing with few and scattered numerical values, we may not need to store these. Then, we can perform One Hot Encoding. For k distinct values, we can transform the feature into a k-dimensional vector with one value of 1 and 0 as the rest values

7. Label Encoding -

Some labels can be words or numbers. Usually, training data is labelled with words to make it readable. Label encoding converts word labels into numbers to let algorithms work on them.

**Machine Learning Algorithms**

There are many types of Machine Learning Algorithms specific to different use cases. As we work with datasets, a machine learning algorithm works in two stages. We usually split the data around 20%-80% between testing and training stages. Under supervised learning, we split a dataset into a training data and test data in Python M Followings are the Algorithms of Python Machine Learning -

**1. Linear Regression**

Linear regression is one of the supervised Machine learning algorithms in Python that observes continuous features and predicts an outcome. Depending on whether it runs on a single variable or on many features, we can call it simple linear regression or multiple linear regression. This is one of the most popular Python ML algorithms and often under-appreciated. It assigns optimali weights to variables to create a line ax+b to predict the output We often u linear regression to estimate real values like a number of calls and costs of houses based on continuous variables. The regression line is

the best line that fits Y-a\*X+b to denote a relationship between independent and dependent variables.

A picture containing indoor, dark, night, night sky

Description automatically generated

**2. Random Forest**

A random forest is an ensemble of decision trees. In order to classify every new object based on its attributes, trees vote for class-cach tree provides a classification. The classification with the most votes wins in the forest. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

![Diagram

Description automatically generated]()

**3. Gradiating Boosting Regressor**

Gradient Boosting algorithm is used to generate an ensemble model by combining the weak learners or weak predictive models. Gradient boosting algorithm can be used to train models for both regression and classification problem. Gradient Boosting Regression algorithm is used to fit the model which predicts the continuous value.Gradient boosting builds an additive mode by using multiple decision trees of fixed size as weak learners or weak predictive models. The parameter, n\_estimators, decides the number of decision trees which will be used in the boosting stages. Gradient boosting differs from AdaBoost in the manner that

decision stumps (one node & two leaves) are used in AdaBoost whereas decision trees of fixed size are used in Gradient Boosting.

**Shape

Description automatically generated**

**VI Dataset**

My dataset is a collection of result of marks obtained by studying for certain number of hours. It is a hybridised dataset as after downloading the dataset from Kaggle I myself added values to make it an appropriate dataset, for this I asked students about their marks and their study hours. By doing so I was able to create a appropriate dataset consisting of samples of students marks and study hours. Given below is how my dataset looks like:-

|  |  |
| --- | --- |
| study\_hours | student\_marks |
| 6.83 | 78.5 |
| 6.56 | 76.74 |
|  | 78.68 |
| 5.67 | 71.82 |
| 8.67 | 84.19 |
| 7.55 | 81.18 |
| 6.67 | 76.99 |
| 8.99 | 85.46 |
| 5.19 | 70.66 |
| 6.75 | 77.82 |
| 6.59 | 75.37 |
| 8.56 | 83.88 |
| 7.75 | 79.5 |
| 7.9 | 80.76 |
| 8.19 | 83.08 |
| 6.55 | 76.03 |
| 6.36 | 76.04 |
| 8.44 | 85.11 |
| 8.41 | 82.5 |
| 7.67 | 80.58 |
| 7.42 | 82.18 |
| 8.16 | 83.36 |
| 5.05 | 70.67 |
| 5.85 | 75.02 |
| 5.45 | 70.96 |
| 7.96 | 83.33 |
| 6.51 | 74.75 |
| 6.73 | 75.65 |
| 5.94 | 74.15 |
| 7.48 | 80.17 |
| 8.13 | 82.27 |
|  | 76.14 |
| 5.4 | 71.1 |
| 8.78 | 84.35 |
| 8.72 | 83.08 |
| 7.1 | 76.76 |
| 7.86 | 81.24 |
| 7.19 | 78.21 |
| 5.62 | 73.08 |
| 7.88 | 83.23 |
| 5.28 | 70.27 |
| 8.92 | 86.41 |
| 5.46 | 71.1 |

**Working:-** Linear regression is used to study the linear relationship between a dependent variable Y (students marks) and one or more independent variables X (study hour).

The dependent variable Y must be continuous, while the independent variables may be either continuous, binary, or categorical. The initial judgment of a possible relationship between two continuous variables should always be made on the basis of a scatter plot (scatter graph). This type of plot will show whether the relationship is linear (figure 1) or nonlinear (figure 2).

Chart, scatter chart

Description automatically generated

Figure 1

A scatter plot showing linear relationship.

Chart, scatter chart

Description automatically generated

Figure 2

A scatter graph showing exponential relationship. In this case, it would not be appropriate to compute a coefficient of determination or a regression line.

Performing a linear regression makes sense only if the relationship is linear. Other methods must be used to study nonlinear relationships. The variable transformations and other, more complex techniques that can be used for this purpose will not be discussed in this article.

As shown below the predicted data is in simpler format. It is showing how our model is predicting the students marks according to the given marks obtained with respect to the study time.

A screenshot of a computer

Description automatically generated with medium confidence

As confusion matrix and accuracy somewhere is used for classification problems. So for my model I used some other techniques to get predicted values with different algorithms and to compare them with each other for that I used ensemble model learning method which is used by Random Forest Regressor. What is ensemble learning it is a technique that actually combines predicted prediction for multiple machine running algorithms.

**Text

Description automatically generated**

another technique we used is Gradient boosting regressor it is one of the most powerful algorithm which can be used for predicting not only continuous target variables but also categorical variable after using gradient boosting regressor I got values which are shown below

**Graphical user interface, text

Description automatically generated**

After that I used baesian ridge another important algorithm. It is used for somewhere poorly distributed data by formulating linear regression so sometimes we have insufficient data or data that is not properly managed so we use baesian ridge for that. After implementing the baesian ridge algorithm in my vsstudio I got the output which is shown below.

Text

Description automatically generated

At last we are using ensemble model accuracy which we got after using various techniques. I am getting accuracy around 0.89 which is 89%.

To get the accuracy for each technique I used so for that I am calculating R2 score, variance score, mean absolute error, mean squared log error.

Actually R2 score is just to evaluate the performance for regression what I have learned is about 02910 your model is cool various code is how far the actual crazy people from the average of predicted values we should not be less than 60% greater than 60% 10% it is how close are printed lines 2 data points means the lower the mean squared log error the higher the accuracy.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Ranom Forest Regressor Report

-> R2 Score: 0.9004973423472099

->mean absolute error: 5.050797201089842

->variance\_score: 0.9006248954903623

-> mean squared log error: 0.05826332430058974

Gradient Boosting Report -> R2 Score: 0.8851455670055234

->mean absolute error: 4.8197961791467545

->variance\_score: 0.8852157300620376

-> mean squared log error: 0.03855770404561921

After printing the R2 score, mean absolute error, variance score and mean squared log error for each algorithm we can see all the algorithms are getting different different values and among them Random forest regressor’s report is the best which is “

Ranom Forest Regressor Report -> R2 Score: 0.9004973423472099

->mean absolute error: 5.050797201089842

->variance\_score: 0.9006248954903623

->mean squared log error: 0.05826332430058974”.

after using different algorithms I chose linear regression and random forest regressor so that I can compare them according to the predicted values.

Chart, bar chart, histogram

Description automatically generated

Here you can see the comparison between them as according to their predicted values.

After that I compared linear regression’s and random forests R2 score and shown below is what I got.

**Chart, histogram

Description automatically generated**

After that I showed comparison between Linear regression and Gradient boosting regressor on the basis of their predicted values and their R2 score. Given below first picture show the comparison on the basis of their predicted values and in the second one it shows the comparison on the basis of their R2 score.

Chart, bar chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

**VII Advantages and Disadvantages**

Advantages of Gradient Boosting in our model are it provides predictive accuracy that cannot be trumped and it can optimize on different loss functions and provides several hyper parameter tuning options that make the function fit very flexible. And also no data pre-processing required - often works great with categorical and numerical values as is. Lets also take advantages of logistic regression its advantages are it is easier to implement, interpret, and very efficient to train it makes no assumption about distribution of glasses in feature space it can easily excel to multiple classes and a natural prognostic view of glass prediction it not only provides a pleasure of how appropriate a predictor but also its direction of association whether positive or negative it is very fast at classifying unknown record accuracy for many simple dataset and it performs well when the dataset linearly separable it can interpret model coefficients as indicators of future important logistic regression is less land overheating but it can over hi dimensional data cells one may consider regularisation L1 and L2 techniques to avoid overfitting these senearios.

Now lets see of the disadvantages of the algo and techniques used to build this model. Gradient Boosting Models will continue improving to minimize all errors. This can overemphasize outliers and cause overfitting. computationally expensive - often require many trees (>1000) which can be time and memory exhaustive. he high flexibility results in many parameters that interact and influence heavily the behavior of the approach (number of iterations, tree depth, regularization parameters, etc.). This requires a large grid search during tuning. Let’s see some of the disadvantages of using linear regression, If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting. The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. Non-linear problems can’t be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios. It is tough to obtain complex relationships using logistic regression. More powerful and compact algorithms such as Neural Networks can easily outperform this algorithm.

**References**

KD Nuggets

<https://www.kdnuggets.com/2022/03/linear-logistic-regression-succinct-explanation.html>

Academia

<https://www.academia.edu/74815512/Measures_of_influence_for_the_functional_linear_model_with_scalar_response>

National Library of Medicine

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992018/>