**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY, BANGALORE**

AN AUTONOMOUS INSTITUTE AFFILIATED TO VTU

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING



**Project Title:**

**MACHINE LEARNING USING LINEAR REGRESSION**

**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

(An Autonomous Institute Affiliated to VTU Belgaum)

Yelahanka, Bangalore-560064



**CERTIFICATE**

This is to certify that the project titled

**MACHINE LEARNING USING LINEAR REGRESSION**

was a bonafide work carried out:

In fulfillment of the requirements for third semester project work under the guidance and supervision of the concerned lecturer during the academic year 2016-17.

**Abstract / Synopsis: -**

Multiple linear regression is a standard technique used in supervised machine learning to establish a relationship between multiple dependent variables and a single continuous dependent variable. This projects implements multiple linear regression algorithms and applies it to multiple problems including housing prices, movie data, city pollution or any other available dataset in any domain.

Table of Contents

[1. INTRODUCTION 6](#_Toc467215342)

[1.1 Purpose 6](#_Toc467215343)

[2. SYSTEM ANALYSIS 7](#_Toc467215344)

[2.1 Existing System 7](#_Toc467215345)

[2.2 Proposed System 7](#_Toc467215346)

[2.2.1 Scope of the Project 7](#_Toc467215347)

[2.2.2 Aim of the project 7](#_Toc467215348)

[2.2.3 Project Modules 8](#_Toc467215349)

[3. REQUIREMENT SPECIFICATIONS 9](#_Toc467215350)

[3.1 Details of Software 9](#_Toc467215351)

[3.2 System Requirements 9](#_Toc467215352)

[3.2.1 Hardware Configuration 9](#_Toc467215353)

[3.2.2 Software Configuration 9](#_Toc467215354)

[4. SYSTEM DESIGN 10](#_Toc467215355)

[4.1 Data Flow Diagram 11](#_Toc467215356)

[4.2 Sequence Diagram 11](#_Toc467215357)

[5. SYSTEM IMPLEMENTATION 12](#_Toc467215358)

[5.1 Modular Description 12](#_Toc467215359)

[6. SAMPLE DATASET 13](#_Toc467215360)

[7. SAMPLE OUTPUT 14](#_Toc467215361)

[8. SYSTEM TESTING 16](#_Toc467215362)

[9. SYSTEM MAINTENANCE 17](#_Toc467215363)

[10. CONCLUSION 18](#_Toc467215364)

[11. BIBLIOGRAPHY 19](#_Toc467215365)

# 1. INTRODUCTION

In recent years, machine learning has found exponentially increasing number of applications for problems such as prediction and classification. Almost all large internet and e-commerce sites use machine learning to understand customers, predict behaviors and tailor user experiences. One of the major types of machine learning systems is supervised learning where an existing large dataset is used to train the system. Multiple linear regression is one algorithm that is used in supervised learning systems.

# 1.1 Purpose

In this project, we implement multiple linear regression algorithms for supervised learning. The design and implementation is completely generalized in order that the program can be applied to any dataset with any number of variables and in any domain.

# 2. SYSTEM ANALYSIS

# 2.1 Existing System

Tasks like predictions on housing prices, pollution levels, movie box-office revenues are typically predicted by experts using heuristics, domain knowledge and intuitive understanding. There were typically no computer based systems which could do these prediction tasks before the advent of machine learning.

# 2.2 Proposed System

The proposed system is a generalized system which takes comma separated datasets with any number of rows and columns in any domain, lets users choose independent and dependent variables, applies multiple linear regression algorithms in order to come up with a best fit equation correlating the variables. It then uses this equation for predicting on new unseen data, similar to the manner in which domain experts would do.

# 2.2.1 Scope of the Project

This project is capable of reading from any CSV data set, processing the information from this data and forming an equation in terms of all the independent variables to obtain the dependent variable’s value. Using the above equation, the system is capable of predicting future values for the dependent variable for new inputs of the independent variables.

This prediction can be used in a vast variety of fields to build business models, make informed decisions when buying or selling products – giving users the same edge as that of an expert.

# 2.2.2 Aim of the project

The aim of this project is to process any CSV data set that is provided to it and predict any dependent variable in the set using one or more different parameters in the given set as the independent variables. The prediction accuracy depends on the relationship the dependent variable holds with the other independent variables. A true linear relationship with all these variables will yield the best results.

# 2.2.3 Project Modules

1. User selection of dependent & independent variables
2. Preprocess dataset
3. Randomized partitioning into training & test datasets
4. Basic matrix operations
5. Get multiple linear regression equation coefficients
6. Apply linear regression equation to test data & user provided data

# 3. REQUIREMENT SPECIFICATIONS

# 3.1 Details of Software

The project has been implemented in C language and shell scripts. It has been developed and tested on Fedora Linux, but is capable of running on any Linux environment.

# 3.2 System Requirements

# 3.2.1 Hardware Configuration

* PROCESSOR: 1GHz or faster x86-class processor
* RAM: 2GB or greater
* Hard disk: 64 GB or greater

# 3.2.2 Software Configuration

* Linux Operating System: Fedora Linux (Version 14 or above)
* For other Linux OS, default command shell must be bash.

# 4. SYSTEM DESIGN

The system is designed as two major modules

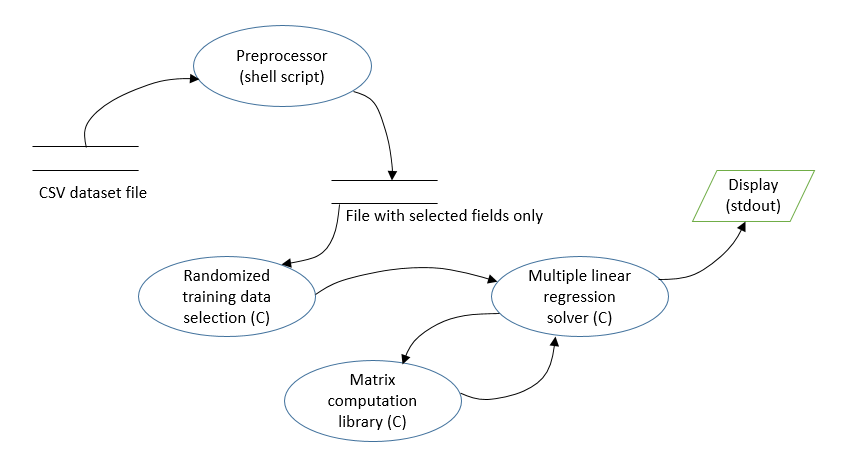
– a bash script which preprocesses the provided dataset to select only the required columns, corresponding to user selected choices of dependent and independent variables;

- a C program which uses extensive n-dimensional matrix operations in order to solve for coefficients of the multiple linear regression equation. All matrix functions are modularly developed and are generalized so as to be able to handle matrices of any size; these are also fully reusable in any other application which might require matrix operations.

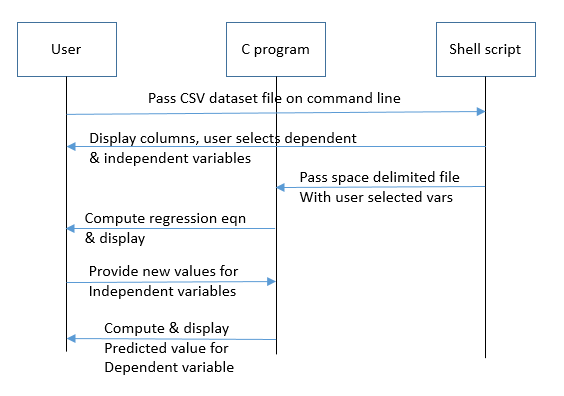
The system also makes no assumptions on the domain, on the number of independent variables, on the size of the dataset or the relative positions of the variables in the dataset; any standard CSV (comma separated values) dataset is acceptable. Also, the system randomly partitions the provided dataset into training data (95% of the dataset) and test data (5% of the dataset).

The application can run on any standard Linux system, though matrix operations start slowing down the system somewhat when independent variables are 10 or more in number. Further improvements in performance for large matrices can be achieved by using advanced algorithms for matrix operations.

# 4.1 Data Flow Diagram



# 4.2 Sequence Diagram



# 5. SYSTEM IMPLEMENTATION

# 5.1 Modular Description

1. User selection of dependent & independent variables

The program allows the user to select the dependent variable and the independent variables. It first checks for headers in the given data set and lists them out for the user to choose from. The bash script accepts the column numbers for the chosen variables.

1. Randomized partitioning into training & test datasets

The C program used the random function to randomly selected 95% of the data set as the training set and the other 5% as the test set. This selection is then stored in a bitmap array that will be then used to divide the data sets into their respective matrices – Trainingset X and Y, Testset X and Y.

1. Basic matrix operations

The multiple regression algorithm consists of various matrix operations such as matrix multiplication, transpose, and inverse.  
To perform each of these operation, matrix functions have been created that work for any number of rows and columns. All these functions are reusable in other applications.

1. Get multiple linear regression equation coefficients

Using the equation from the algorithm, which involves the matrix functions, the coefficient matrix is generated. This coefficient matrix is used to create an equation in independent variables to find the dependent variable.

1. Apply linear regression equation to test data & user provided data

Once the equation has been generated, it can be used on any data – either the remaining 5% of the original dataset or on user provided inputs to predict the dependent variable’s value for new data.

# SAMPLE DATASET

Shown below is a subset of a dataset showing pollution data in various parts of the state. Any data in CSV format, with columns representing variables and rows representing individual data measurements are acceptable to this system.



# SAMPLE OUTPUT

The header fields are these:

1 : id

2 : date

3 : price

4 : bedrooms

5 : bathrooms

6 : sqft\_living

7 : sqft\_lot

8 : floors

9 : waterfront

10 : view

11 : condition

12 : grade

13 : sqft\_above

14 : sqft\_basement

15 : yr\_built

16 : yr\_renovated

17 : zipcode

18 : lat

19 : long

20 : sqft\_living15

21 : sqft\_lot15

Please enter the field number for the field that is to be predicted:

3

Enter field numbers of independent variables one by one, use 0 or invalid number to end:

4

5

6

7

0

Std Error of Regression is: 244601.044210

The multiple linear regression equation is:

price = 77408.829304 + -58917.527243 \* bedrooms + 6447.793422 \* bathrooms + 314.073464 \* sqft\_living + -0.384539 \* sqft\_lot

To calculate price :

Enter bedrooms

2

Enter bathrooms

3

Enter sqft\_living

4556

Enter sqft\_lot

5667

The price for the given values is: 1407656.673427

# SYSTEM TESTING

The system was tested extensively with a number of different datasets from different domains including housing prices, movie gross revenues, city air pollution, student performance etc. Testing was also done for boundary conditions (e.g. only 1 independent variable, large number of independent variables). Negative testing was also done (e.g. attempting to repeat variables, out of range values, invalid values etc.). Error messages were also checked for, to ensure that any user errors in input are notified and users asked to correct). During the course of development, each function was independently tested by exercising it through a range of allowable and illegal values.

# SYSTEM MAINTENANCE

The system has been designed and implemented so as to be applicable for any domain and any size of data (constrained however by memory and processing power available on target system). A number of functions (such as matrix, bitmap array, randomization) can be easily reused or adapted in other applications. Code is also documented to ensure readability. Further future enhancements to this system could include automatic mapping of categorical alphanumeric data into numerical categories so that these too can be used as variables for regression analysis.

# CONCLUSION

This project successfully implemented a multiple linear regression equation solver which is generalized enough to run on any available dataset. The software was successfully tested against a number of publicly available datasets such as housing prices datasets, movie datasets, student grade datasets etc. Multiple linear regression is a key tool in machine learning and data science, used extensively by data scientists worldwide. The algorithm used is a standard multiple linear regression solver, which is part of the standard toolbox in environments such as R and Python which are used extensively by data scientists. The project also leverages different tools according to their strengths; shell scripting for processing and extracting from large CSV dataset files; C program for computationally intensive matrix operations, random selection of data and finding coefficients of multiple linear regression equations.

# BIBLIOGRAPHY

Multiple Linear Regression algorithm: <http://www.originlab.com/doc/Origin-Help/Multi-Regression-Algorithm>

Machine Learning algorithms: <https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-choice/>

Regression Analysis: <https://en.wikipedia.org/wiki/Regression_analysis>

Some datasets for machine learning: <https://archive.ics.uci.edu/ml/datasets.html>