EDResearch

Updates

1. General all round updates
2. Credit Score

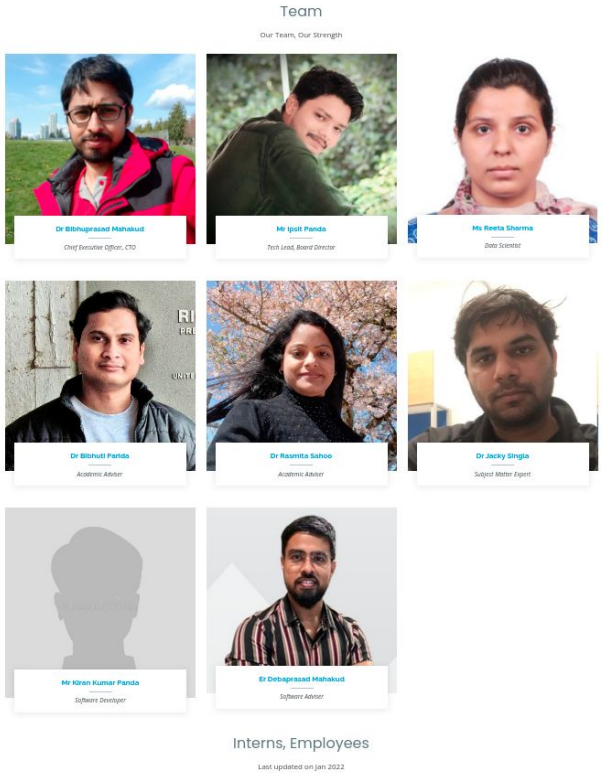
1st May, 2022 | B.Mahakud

General Company Updates

– – –

- New company diracAI pvt ltd name **reserved!**
- Company registration in progress. Will take few days.
- diracAI.com Web development progress.
- Our app website and company website will be different from each other.
- App website: edresearch.co.in updated. Team/Career sections updated. Employee accounts added.
- Latest codes are pushed to production.

General Company Updates



- Both edresearch.co.in(or may be not) and diracAI.com will have team section.
- diracAI.com will be used for operations, employees and investors, interns
- edR app will be a product of diracAI pvt ltd.
- diracAI.com will be designed keeping in mind the needs of investors, employees and govt rules and regulations, privacy policy, terms and conditions
- edresearch.co.in will be customer oriented.

General Company Updates

- Dr. B. Mahakud: Back-end, react app front end, app-website, Credit score development
- Er. Sutar: React Video player development
- Dr. Singla and Ms. Reeta: Credit score
- Mr. Ipsit Panda and Mr. Kiran: Android App
- Dr. B. Parida : Credit Score deployment and development through interns
- Er. Debaprasad Mahakud: diracAI.com web development

```
We will discuss how credit score work could be divided among the participants.
```

Problems : Teachers

Teacher needs to monitor

- Student's progress
- Want to know causes of BAD performance in exams
- Teacher needs to ensure if the students are completing their responsibilities in time

Students' responsibilities:

1. Attend class regularly
2. Finish homework
3. Pay attention in class
4. Do self study
5. Finish prerequisites of each class and so on

How can a teacher keep track of this for 100s of students??

Answer: Credit Score

Online Learning

Blended Learning

Problems: Private Institutes, Univs, Digital Univs, IITs

- Private institutes are very concerned about student monitoring for large number of students
- No Single App exists which does efficient monitoring of activities, student understanding and progress.

- From office monitor student progress
- Know which courses are going well
- Which colleges are completing their responsibilities in time.
- If the courses are being taught in time.
- If both teachers and students are doing their work properly.

Why a single number is required for monitoring ?

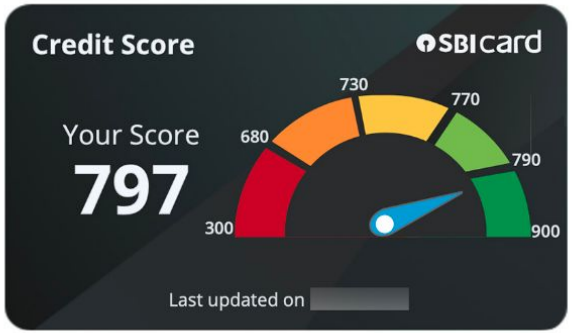
- There are applications (moodle, teachmint) which gives several information about student activities like attendance.
- But those data becomes useless if the number of students are very large.
- It is very difficult to look at 10 different graphs of 1000 students to conclude anything!!

We need a single number which will reflect all the parameters i.e. credit score

- Not attending class => Low credit score
- Not completing homework => Low Credit score
- Not alert in class => Low Credit score
- Student is asking lot of questions => will enhance credit score
- Not understanding in class => Low credit score

Similar examples in Banking Industry : Credit score

- A single number is calculated from various inputs like financial background, salary, previous loan defaults etc. to form a number called credit score
- Banks use this credit score to decide if they want to give loans to the individual.
- If you have a good paying job, your credit score will be high.
- If you are living in a politically unstable country , credit score will be less.



How Bank Credit Score is calculated?

– – –

Credit Score = c1*Salary + c2 * personal Assets + c3 * bank balance + ...

A normalized number is calculated between 0 to 1000.

Credit Score tells how likely you will be paying back the loan amount to the bank.

Our case is similar with subtle difference !!

Bank Credit Score	edR Credit Score
It tells how likely an individual is expected to pay back loan.	**First Priority:** It will tell if the responsibilities are completed **Second Priority:** It will tell how likely a student in expected to perform in exam.
Uses a complex ML algorithms to train the data. Relative weights may not be retrievable.	We need to know the relative weights of input parameters. This will help us to provide simplified meaning to institutes directors /managers.
They can use Deep Neural network.	Most likely we cannot but deep neural network can be used to compare performance.

What input parameters to be considered?

1.) Attendance in Class
2.) Alertness in Class
3.) Questions Asked during Class
4.) Understanding of topics taught during the class
5.) Completion of prerequisites for a class
6.) Home-work of a class
7.) Difficulty level set in homework
8.) Time taken to complete homeworks
9.) Set difficult level in the questions from book.
10.) Reading and understanding after class
11.) Questions /Doubts from reading text book.
12.) Motivation towards the subject

10.) Psychology
11.) Ambition of the student
12.) Financial Background
13.) Study room atmosphere / hostel
14.) Previous Exam performance , previous class performance
15.) If external questions asked either by teacher or student, to be rated by teacher
16.) Time spent in the edrApp
17.) Reading of external notes
18.) Audio issues during the class
19.) GPS location Data from mobile app

Software, Programming Language, Techniques

- Python , PyROOT
- Software for Statistics: ROOT
- ML Techniques : BDT, xgboost, DNN, Keras, Tensorflow
- Bayesian Likelihood

https://root.cern/install/

Download a pre-compiled binary distribution

We distribute pre-compiled ROOT for several major Linux distributions as well as MacOS and (as a beta) Windows. The steps to install a pre-compiled binary are simple:

1. Install all required dependencies with the system package manager
2. Download the release for the desired platform and ROOT version
3. Unpack the archive
4. Add the ROOT libraries and executables to your environment by sourcing the appropriate `thisroot.*` script. These setup scripts can be found in the ROOT binary release, in the `bin` directory.

For example, on Ubuntu 20, a user could execute the following bash commands to install ROOT v6.24/02, after installing all required dependencies:

```
$ wget https://root.cern/download/root_v6.24.02.Linux-ubuntu20-x86_64-gcc9.3.tar.gz
$ tar -xzvf root_v6.24.02.Linux-ubuntu20-x86_64-gcc9.3.tar.gz
$ source root/bin/thisroot.sh # also available: thisroot.{csh,fish,bat}
```

Input distributions: Attendance

- Attendance : Number of students coming to class on a given day.
- What distribution will it follow ??

Poisson? Gaussian? Or something more complex??

How will we start thinking ??

Recap: Binomial distribution

Coin toss experiment:

P: Probability of success in a single toss ~ 0.5 (identical coins)

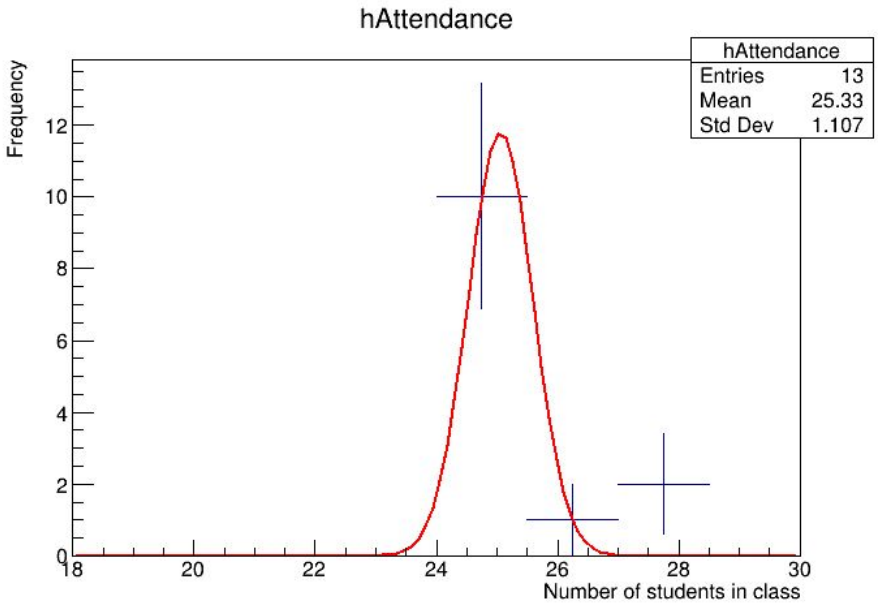
P finite, Trials are large=> Poisson

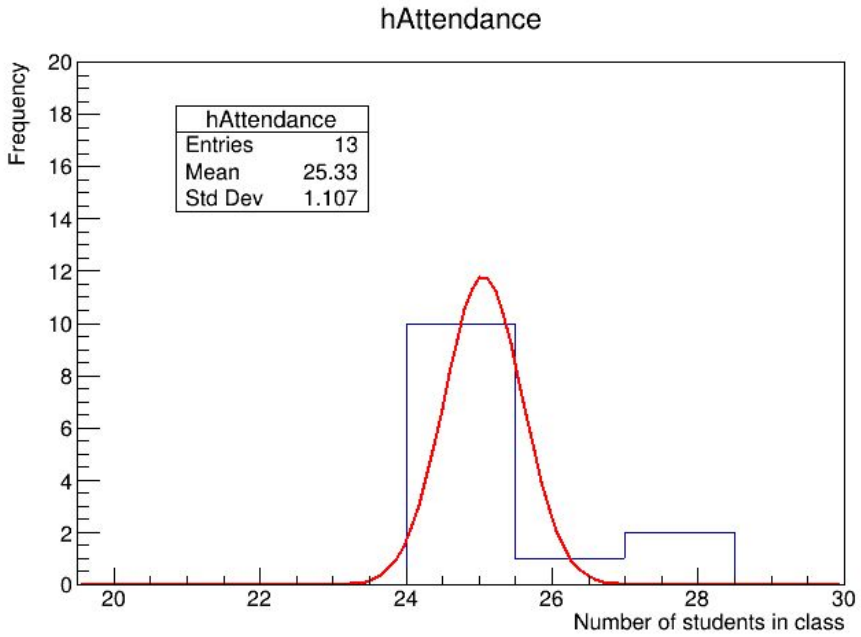
$$f(k;\lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

Student coming to class:

P: Probability of coming to class on a single day. varies ~0.3 -1 (identical students ~ not identical but lets assume identical i.e. p~ to fix number. Same for all)

Attendance: Real Data(From Dr. Parida)

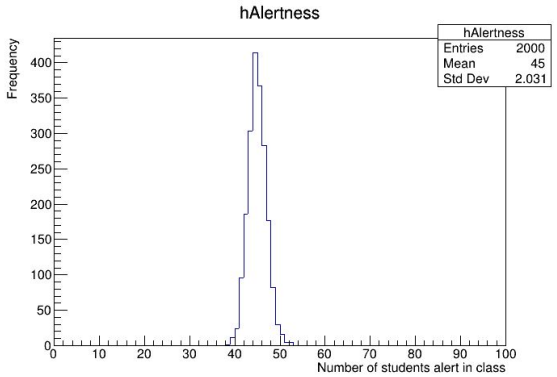




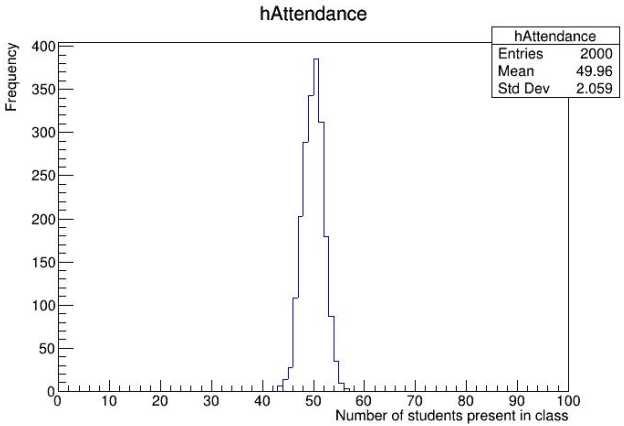
Consistent with gaussian!! Very little data. We can confirm when we have more data

Input Parameter distributions

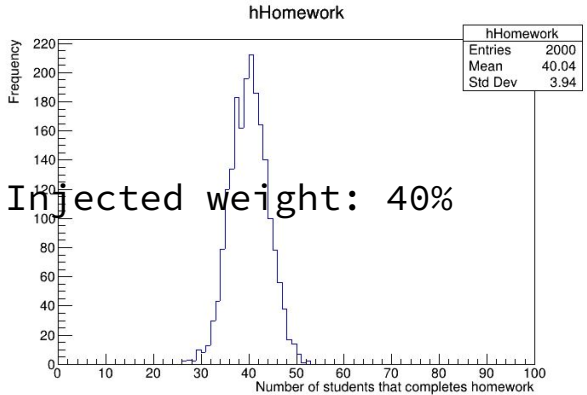
Injected Weight: 40%



Injected Weight: 20%



Injected weight: 40%

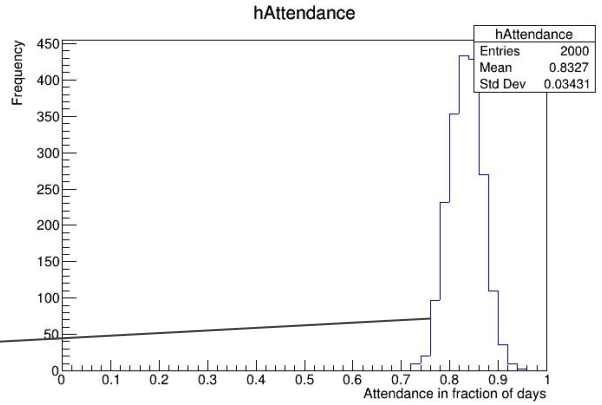


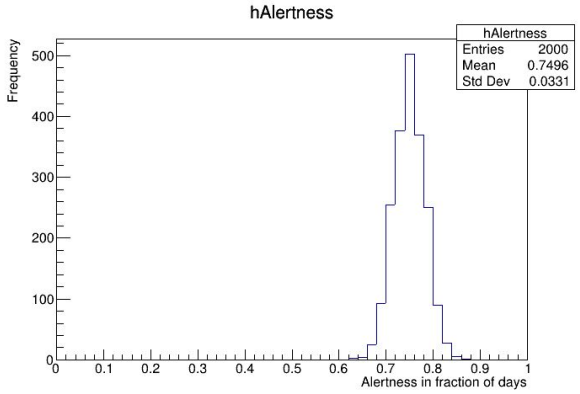
Prob(credit score)= (0.4*Alertness+ 0.2*Attendance+0.4*Homework) + constant

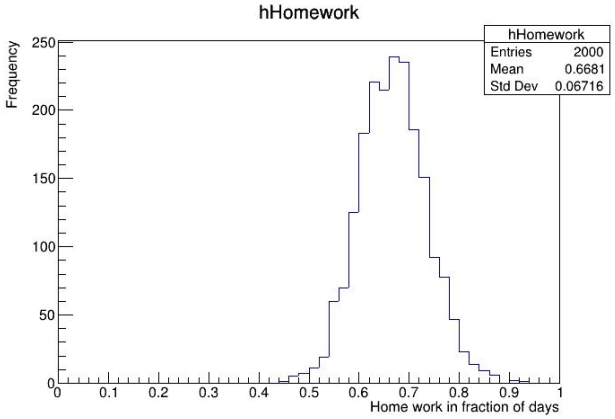
Normalized Distributions

- These are normalized distributions = Previous distribution /strength of the class

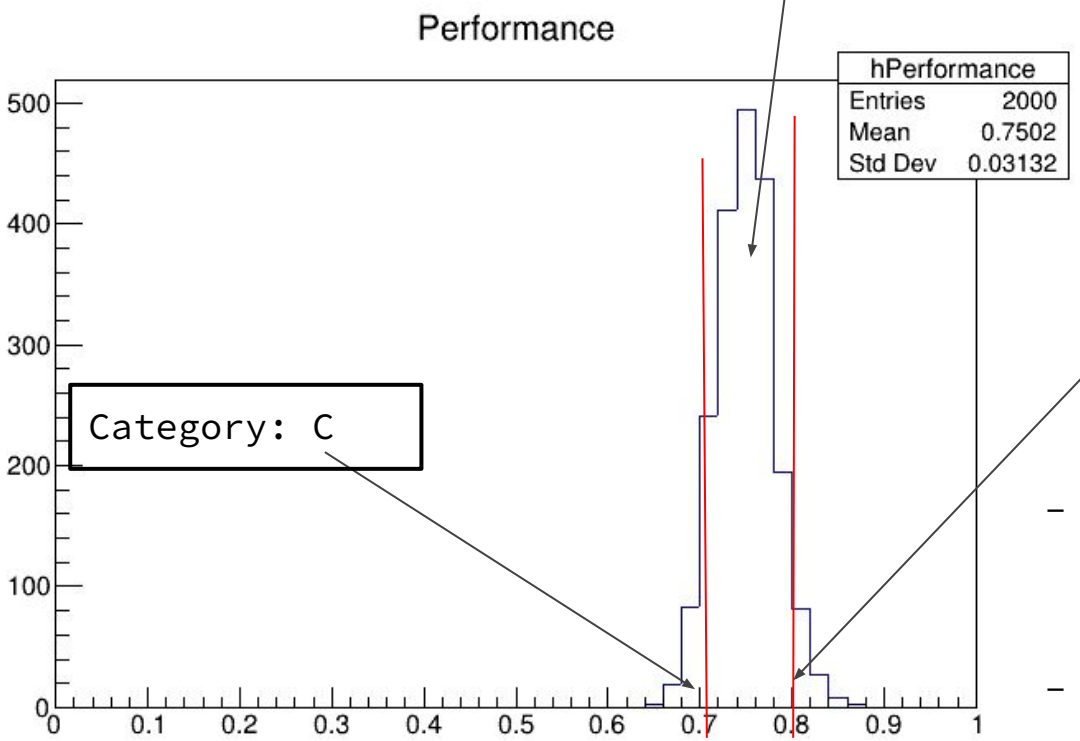
X-axis: Maximum students have attendance close to 83% throughout the year.







Exam Performance



Category:B

Category: A

- **Linear regression**
- **Logistic regression**

Category: C

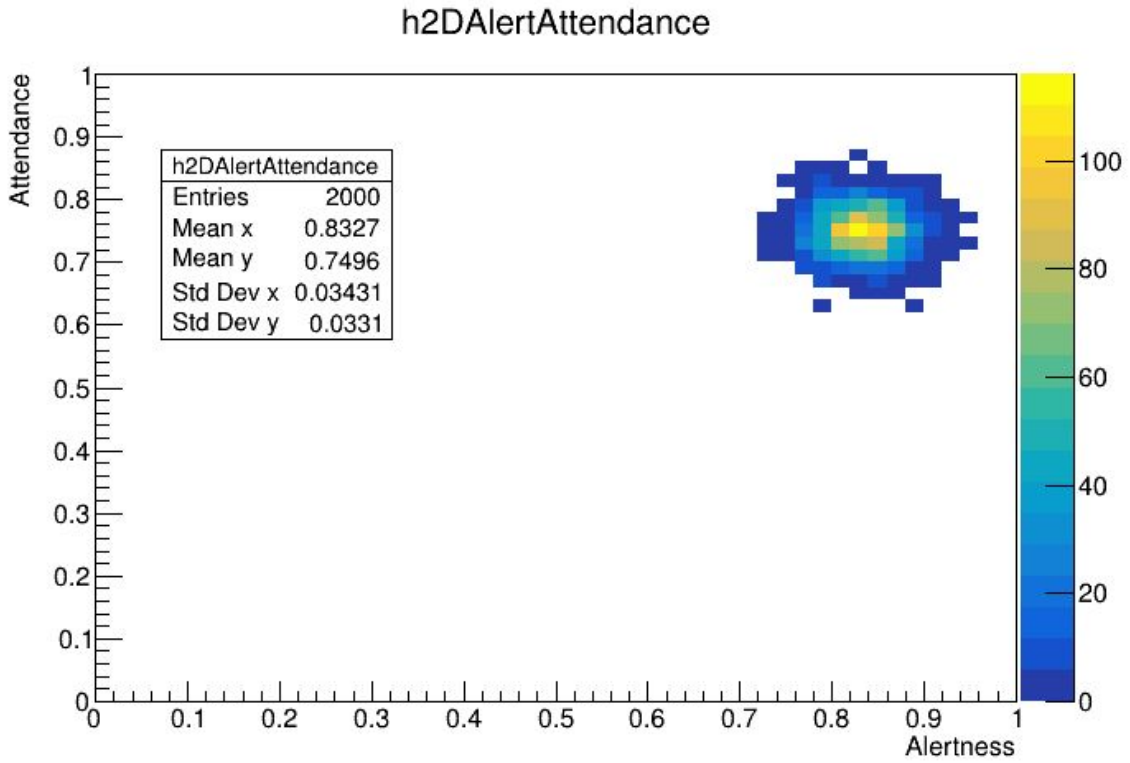
- If performance is available in terms of grades => Logistic regression
- Performance in percentages => multi-linear regression

Simulated Training Data : 2000 points

Alertness	Attendance	Homework	Performance
0.78	0.62	0.62	0.70~70% (B)
0.82	0.55	0.55	0.69~69% (C)
0.85	0.71	0.71	0.76~76% (B)
0.83	0.61	0.61	0.73~73% (B)
0.84	0.58	0.58	0.72~72% (B)
0.95	0.80	0.80	0.85~85% (A)
-	-	-	continue…

Simulated Data

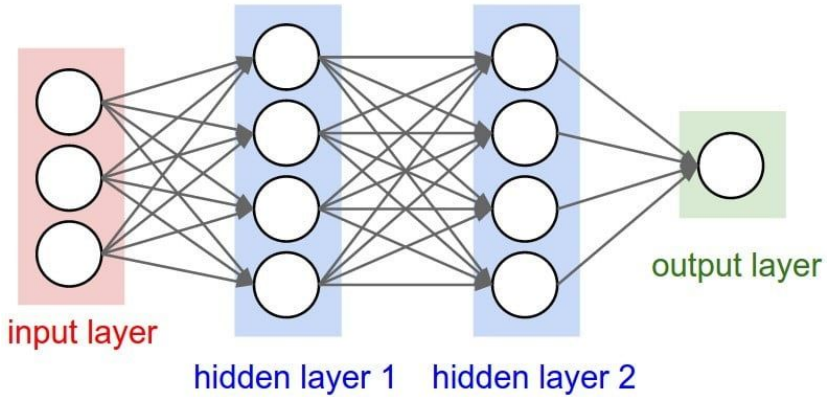
Correlation: 2D Scatter plot



- But in real data the distributions could get more complex with complex correlations
- When correlations becomes complex, we need complex machine learning techniques like deep neural network.

What Machine Learning Technique we should use??

- Multi - layer deep neural network is not a good option?? Why?
- Because it is difficult to develop intuition about the working and covey to layman.(i.e. Teachers, directors)
- But DNN can be used to set bench mark performances



It is difficult to translate the hidden layer weights to an interpretation to layman or even to a learned man!!

Our Best bet will be: Regression (old and classic, and still used by many for this reason in industry)

What if regression model have low efficiency ?

- Regression as expected cannot model very complex correlations among input variables.
- This could lead to low efficacy if the number of variables will increase.

Possible Solution:
Decorrelate the input variables before applying regression

Conventions for ML work

Coursera conventions: by Andrew Ng

$$X1(3 \times 1) = \begin{vmatrix} 0.78 \\ 0.62 \\ 0.62 \end{vmatrix} \quad \begin{matrix} \text{(attendance)} \\ \text{(alertness)} \\ \text{(homework)} \end{matrix}$$

Single training input

X = Training Set (nVars X Num Training Examples) = [X1 X2 . . .]
Y = Output

nVars = n_x
Num Training Examples = m
X ~ (n_x, m) , Y = (Y1, Y2. . .m times) ~ (1xm)

Regression

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_n x_n$$

$$h_\theta(x) = \begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_n \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} = \theta^T x$$

Where x1,x3 … are attendance , alertness .. etc.

Cost function:

$$J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \dots, \theta_n)$$

How to calculate uncertainty on Credit score?