Logistic regression

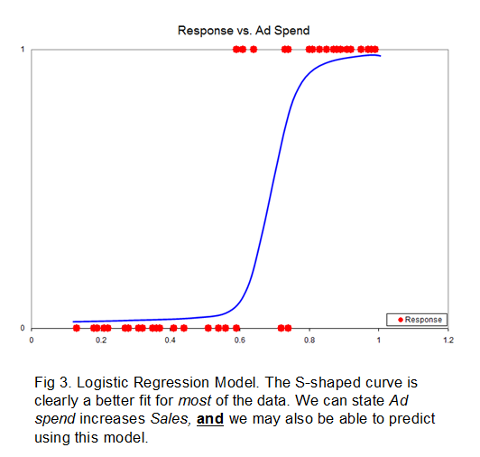
Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome,

Where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logic function.

logit model/Logistic Regression is a [regression](https://en.wikipedia.org/wiki/Regression_analysis) model where the [dependent variable (DV)](https://en.wikipedia.org/wiki/Dependent_and_independent_variables) is [categorical](https://en.wikipedia.org/wiki/Categorical_variable).

**Binary classification**

Binary classification is performing the task of classifying the binary targets with the use of [supervised classification algorithms](https://dataaspirant.com/2014/09/19/supervised-and-unsupervised-learning/). The binary target means having only 2 targets values/classes. To get the clear picture about the binary classification lets looks at the below binary classification problems.



### Example: Probability of Issuing a credit card versus Salary.

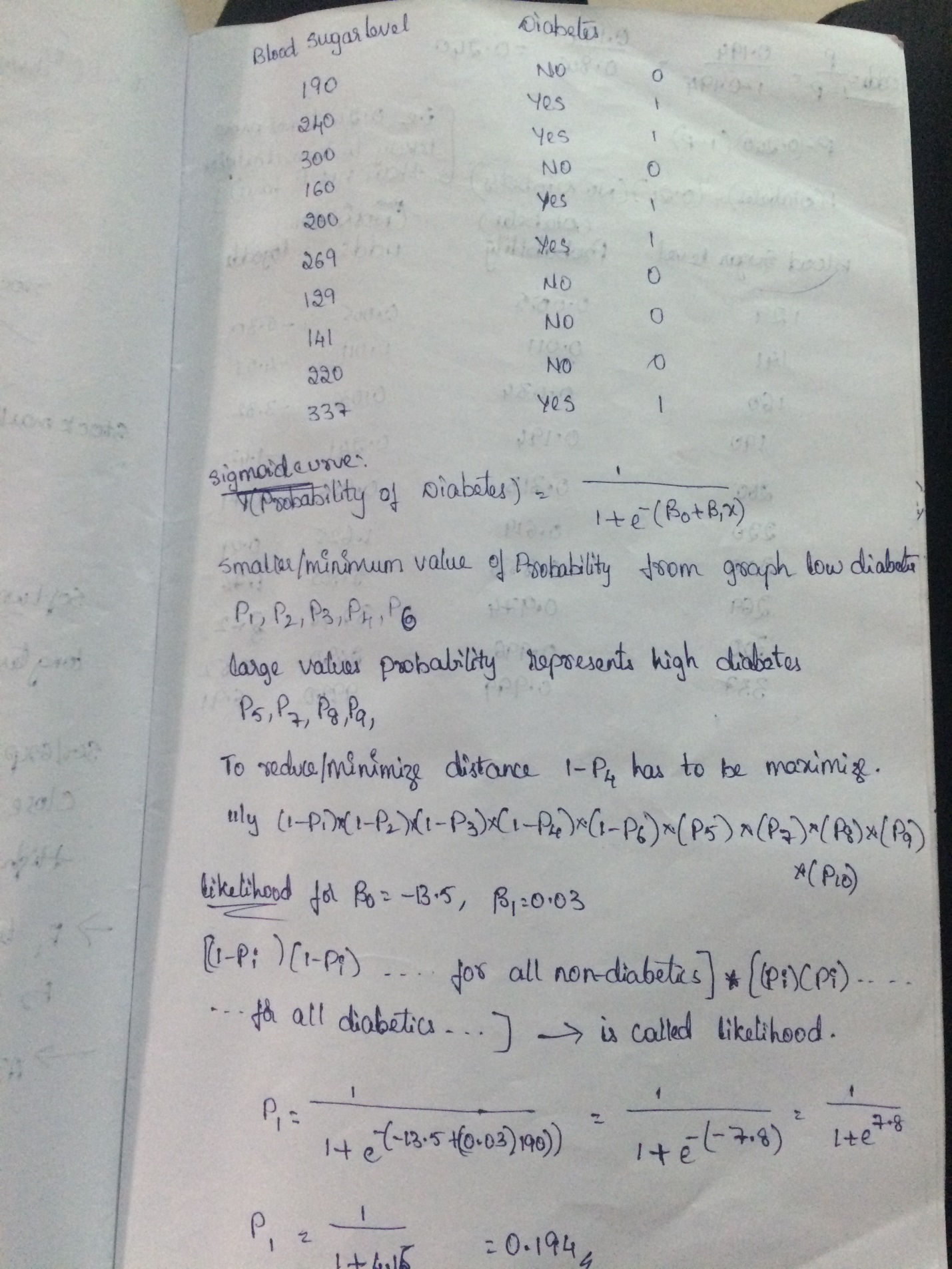
### A group of 20 People applied for a credit card. How does the salary20000 affect the probability that the person will get the credit card?

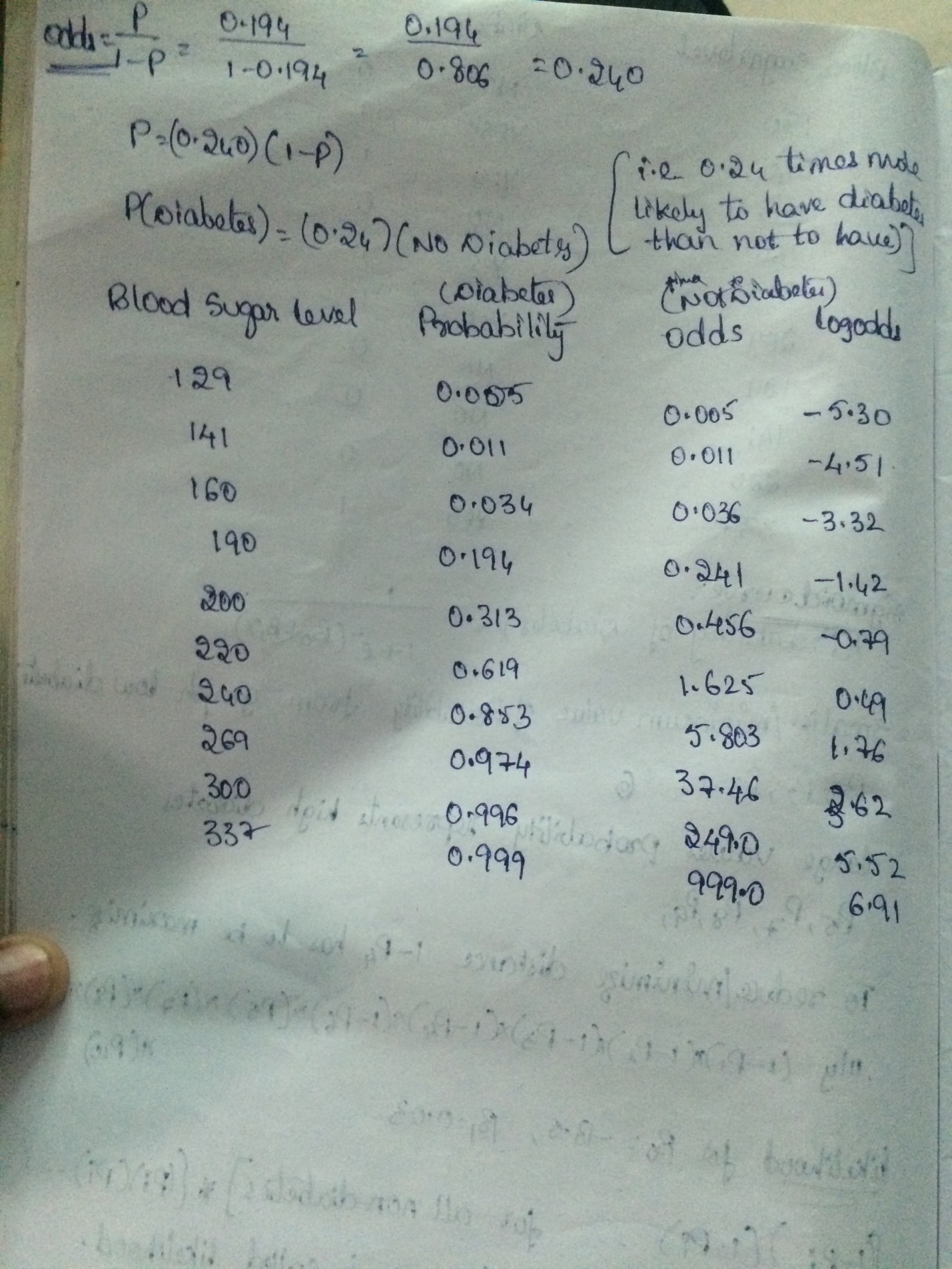
### The reason for using logistic regression for this problem is that the values of the dependent variable, Yes and No, while represented by "1" and "0", are not [cardinal numbers](https://en.wikipedia.org/wiki/Cardinal_number). If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) could be used.

The table shows the Salary of each person and whether they issued credit card YES (1) or NO (0).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age** | **salary** | **marital** | **loan** | **Y** |
| 58 | 100000 | married | no | 0 |
| 44 | 60000 | single | no | 0 |
| 33 | 120000 | married | yes | 0 |
| 47 | 20000 | married | no | 0 |
| 33 | 0 | single | no | 0 |
| 35 | 100000 | married | no | 1 |
| 28 | 100000 | single | yes | 0 |
| 42 | 120000 | divorced | no | 0 |
| 58 | 55000 | married | no | 1 |
| 43 | 60000 | single | no | 0 |
| 41 | 50000 | divorced | no | 0 |
| 29 | 50000 | single | no | 1 |
| 53 | 60000 | married | no | 0 |
| 58 | 60000 | married | no | 0 |
| 57 | 70000 | married | no | 0 |
| 51 | 55000 | married | no | 0 |
| 45 | 50000 | single | no | 0 |
| 57 | 20000 | married | no | 0 |
| 60 | 55000 | married | no | 0 |
| 33 | 70000 | married | no | 0 |
| 28 | 20000 | married | yes | 0 |
| 29 | 20000 | married | no | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| Summary statistics: |  |  |  |
|  |  |  |  |
| Variable | Categories | Frequencies | % |
| y | 0 | 18 | 81.818 |
|  | 1 | 4 | 18.182 |





Application:

1. Prediction - logistic regression allows us to give probability estimates relevant to making predictions.
   * Medicine, assessing the risk of cardio-vascular disease based on current health and habits
   * Framingham risk equation - derived from historical data using logistic regression
   * [Online Tool](https://www.cvdriskchecksecure.com/FraminghamRiskScore.aspx)
2. Isolating the effects of single variables
   * Risk factors such as smoking, alcohol consumption
   * early development of logistic regression motivated by smoking studies
   * Beneficial factors such exercise, good dietary habits
3. Understanding the “interplay” between several factors
   * interactions
   * mediation

**UseCases-**

1. Finding the difference between approval of credit card
2. Finding obese person with adiposity prone to heart disease
3. Finding Product Price & Sales
4. Finding Age & Mortality
5. Finding temperature vs. Number of cones sold at ice cream store
6. Finding Population vs Food consumption
7. Finding quantity with yield
8. Determining the chances to win cricket match .
9. Determining the chances of getting Jobs after Completing Graduation.
10. Speed and distance relationship
11. Finding rate of growth of the economy of a Institution

**PYTHON CODE**

import numpy as np

import pandas as pd

from matplotlib import pyplot as plt

from sklearn.linear\_model import LogisticRegression

from sklearn.cross\_validation import train\_test\_split

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import accuracy\_score

creditdata=pd.read\_csv("C:\\Users\\Rama\\Desktop\\Bank.csv")

creditdata

print(creditdata.head())

print(creditdata.describe())

print(creditdata.corr())

features=creditdata[["salary","Age","y"]]

targetVariables=creditdata.y

featureTrain,featureTest,targetTrain,targetTest=train\_test\_split(features,targetVariables,test\_size=0.3)

model=LogisticRegression()

fittedModel=model.fit(featureTrain,targetTrain)

predictions=fittedModel.predict(featureTest)

print(confusion\_matrix(targetTest,predictions))

Output:

runfile('C:/Users/Rama/Desktop/Python Project/Linear regression/Logistic regression/Logistic regression Python.py', wdir='C:/Users/Rama/Desktop/Python Project/Linear regression/Logistic regression')

Age salary marital loan y

0 58 100000 married no 0

1 44 60000 single no 0

2 33 120000 married yes 0

3 47 20000 married no 0

4 33 0 single no 0

Age salary y

count 22.000000 22.000000 22.000000

mean 43.727273 59772.727273 0.181818

std 11.423375 32786.367402 0.394771

min 28.000000 0.000000 0.000000

25% 33.000000 50000.000000 0.000000

50% 43.500000 57500.000000 0.000000

75% 56.000000 70000.000000 0.000000

max 60.000000 120000.000000 1.000000

Age salary y

Age 1.000000 0.005548 -0.252467

salary 0.005548 1.000000 -0.051842

y -0.252467 -0.051842 1.000000

[[6 0]

[1 0]]