Script for Presentation – 3/4/2021

Good Morning. My Name is Radha and I am part of Team 8. We are presenting Lambda and associated functions.

Request you to park all your questions till the end of both ITP and NPV presentations.

**Dinesh : Move Slide to Lamda**

1. A lambda function is a one line anonymous function, anonymous because it has no name.

2. A lambda function can take any number of arguments, but can only have one expression.

3. The way the lambda function works is the expression is executed using the arguments we pass and the result is returned.

4. If we assign a identifier to the function object returned by lambda, it behaves similar to a user defined function

5. In upper right corner we have a simple lambda function example, first one returns square of a number and the second a cube.

**Dinesh : Move Slide to Map**

Map()

1. The map() function executes a specified function for each item in an iterable. An Iterable is an object, which one can iterate over. e.g List

2. The way I imagine map function is we put some input into a machine and pull out a final transformed output.

3. The way it works is map() applies tranformation to each item in an iterable.

4. In the upper right corner we can List as an iterable using map gives a new list which can be squares or cubes extra.

5. Lower right corner we have map() and map() with lambda.

**Gopal will present**

**Dinesh : Move Slide to Filter**

Filter()

The filter() method filters the given sequence with the help of a function that tests each element in the sequence to be true or not.

function: function that tests if each element of a sequence true or not.

sequence: sequence which needs to be filtered, it can be sets, lists, tuples, or containers of any iterators.

Returns: returns an iterator that is already filtered.

**Dinesh : Move Slide to Reduce**

 reduce()

It is a very useful feature when we need to apply a function to an iterable and reduce it to a single cumulative value.

As an example , we have a stack match box, every reduce iteration pulls out the match sticks from the match box and adds them to the next box and moves on. Finally we are left final count of match sticks.

In this process we are converting large dataset into a single value to perform further operations

**Dinesh : Move Slide to Lamda with other functions**

**Analysis of Diabetes dataset using Visual Expression**

**Amar starts here -**

Good morning everyone. Today we explore Visual Exploration and Analysis of Diabetes Dataset.

To start with, our dataset has a total of 768 rows. The data refers to females ages between 21 to 81 years.

There are total nine attributes can be further classified as-

Six attributes describe the result of physical / verbal examination

1) Number of times pregnant

2) Diastolic blood pressure

3) Triceps skin fold thickness

4) Body mass index

5) Diabetes pedigree function - a function which scores probability of getting diabetes based on family history

6) Age in years

B) Two attributes are result of chemical examination

7) Plasma glucose concentration over 2 hours in an oral glucose tolerance test

8) 2-Hour serum insulin

C) Resultant Target Variable

9) Class variable - where 0 stands for non-diabetic, and 1 stands for diabetic

The Objective of this data analysis is to find the significant factors in the first 8 variables which influence the 9th variable that is diabetes or not, through data analysis and visualization techniques using Python.

Diabetic Dataset and Jupyter program is hosted in Github to enable concurrent and collabrative working within the team.

First, we are reading data from github repository.

Next, we are checking info of data frame.

We have 9 columns, 768 rows and data types of all columns are displayed below.

Next, we are checking for NaN in the data frame. WE can see here there is no NaN Data frame.

Finally, we are checking for duplicate rows in the data frame and there is no duplicate rows as well.

Next, I will invite Dinesh to explain about Data Cleaning and exploration.

**Dinesh starts here –**

Hi this is Dinesh, from Team 8 , we started cleaning the data by generating descriptive statics of mean ,SD, min max and quartile deviation of variables. While doing this we notices minimum value for many variables are 0 ,as biological parameters like glucose , BP, skin thickness , insulin and BMI cannot have zero values , we can infer that null values have been coded as 0

Checking for 0 count , we see Glucose , Blood Pressure and BMI have less than 5% 0 values , hence dropping them would not significantly impact the sample size.

However Insulin with 374 and Skin thickness 227 in a data set 768 is almost 50% , hence we need to impute them with suitable Mean or Median

We can see from the box plot for insulin against Class variable that there are many outliers which are skewing the mean by around 30 points and there is also a significant 70 point difference between the mean and median . we decided to replace the zero values of ND with median 102.5 and D with 169.5

In the Box plot for skin thickness and class variable we see only marginal difference between mean and median hence we decided to replace 0 with their respective Mean

This is the code to impute 0 with suitable Mean Median for insulin and skin thickness . we dropped the 0 value rows for low impact Glucose , BP and BMI.

We also renamed the long variable names with shorter Alias to help visualization through graphs

You can see we have been able to successfully impute the 0 values with only 6% sample size reduction.

Moving on to Data exploration we started with Univariate analysis . Pie Plot of resultant class variable shows that 65% of sample population are non diabetic

we plotted the Histogram and density graph for other variables.

Insulin and Pedigree are right skwed with long tail

Glucose and results are left skewed and rest of the variables are more or less centered

Interestingly no bi modal series were observed ,so there could be no proper demarcation between groups , let us further check using Bi Variate analysis

We used pair plot to compare inter relation between variables

The diagonal shows the main effect of the variable on itself in sub group of D /ND. For example the density plot of 2nd row glucose shows distinct peak point for D /ND , but there is also large over lap between the groups .

The interaction effect of glucose with other variable shows some distinct grouping in scatter plots. Let us further explore using pearson co relation coefficient and heat maps

If we look horizontally in the last low , we can see resultant variable is most co related to Glucose followed by others . Incidentally Glucose is also equally co related to insulin

Above is the ranking of resultant co relation , as seen from the heat map even though insulin is 2nd after glucose only one of this variable is recommended for use in model to avoid any inflation variation.

Let us further deep dive into Data Visualization and Analysis.

**Subhankar starts here -**

Data Visualisation - Diabetes

First lets take a look at Density and Distribution plotting using Violin plot and Dist plot to get data-distribution of:

Serum Insulin

Plasma Glucose

Result:

Insulin: Higher Insulin level indicates the presence of higher amount of blood sugar. For majority of Diabetic patients(violin plot - mean as well as density), 2 hr-Insulin level is higher(@170-180mu), while for Non-Diabetic patients it is around 100. The distplot shows that as insulin level increases, so does the count of the diabetic people

Plasma Glucose Level: Diabetes is caused when the blood sugar level is too high. There is a clear difference in the mean values of PGC level between Diabetic(142-145) and Non-Diabetic(108-110) sample. The distplot clearly depicts that healthier sample is at lower glucose levels, while with increase in glucose level, the no. of diabetic patients increase

Density and Distribution plotting using Violin plot and Dist plot to get data-distribution of:

BMI

Triceps Skin Thickness

Result:

BMI: Below 18.5 Underweight 18.5 – 24.9 Normal or Healthy Weight 25.0 – 29.9 Overweight 30.0 and Above Obese . The more weight you have, the more your body becomes resistant to insulin. the mean of non-diabetic patients is 30, which is borderlining Obese, while the majority of the sample is around 25.0. While for the Diabetic sample, the mean as well as the majority is in the range of 34-37. We can infer that increase in BMI lead to higher chances of being diabetic

Tricep Skin Thickness: Diabetes causes a condition leading to increase in the thickness of the skin. The dist plot clearly shows that majority of diabetic patients are having thicker skin at triceps as compared to healthier sample.

Density and Distribution plotting using Violin plot to get data-distribution of:

Age

Pregnancy Count

Result:

Age: The data clearly shows that at lower age of 20-25(First quartile to mean) most of the population is healthy and without diabetes. The distribution graph gives a clear indication, that with increase in age, the majority of the population is diabetic

No of Times Pregnant: The violin plot for pregnancies as well as the dist plot shows it clearly that the highest density of the both diabetic and non-diabetic sample can lie in the same range. This leaves this information as non-deterministic wrt chances of being diabetic

Density and Distribution plotting using Violin plot to get data-distribution of:

Pedigree

Blood Pressure

Result:

Diabetic Pedigree: This is basically the analysis of family genealogy to determine if it can be used to determine that someone is diabetic or not. Although the number having diabetes increase with the increae in the value of the pedigree, still the data at the max density is overlapping in both the plots. This is not a very clear indicator of determining if somebody is diabetic or not.

Blood Pressure: This is usually a very clear indicator, as diabetic patients have caused damage to their blood vessels due to increase in blood glucose level. But in the given dataset, the count of non-diabetic individuals with normal blood pressure of 80 is much high(388), as compared to the count of diabetic individuals with blood pressure above 100(9). This has caused the data to not clearly emphasize the impact of blood pressure on Diabetes

From the above analysis, we can conclude the following: Deterministic Attributes

Data provided in Insulin is the best indicator for detecting diabetes condition

Plasma Glucose Level are the most clear indicators of detecting if a patient is Diabetic or not.

BMI Level is is the next best potential indicator and with higher weight conditions, the chances of being diabetic is much higher.

AGE

Tricep Skin Thickness

Non Deterministic Attributes

Pregnancy

Diabetic Pedigree

Blood Pressure

Also, our analysis are consistent with the correlation rank that we created for every column wrt Class variable.