Name Chopra Aman, Ranjan Nidhi Date: 4th December 2022

NYU NetID: ac8511, nr2387

Section: CSCI-GA.2433-001

**Project Part 3**

We Aman Chopra and Nidhi Ranjan, hereby certify by submitting this project Part 1 that all the efforts put into this part are our own. We have referred the project support materials.

**Total in points** (100 points total): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Professor’s Comments:**

**A picture containing arrow

Description automatically generated**

**Affirmation of my Independent Effort:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**EDA Physical Database Design:**

**Changes made in part 2 of the Project:**  
We added a field called Amount in the Transaction Table as it was missing.  
We changed the Type of ‘Recurring’ Field in the Policy Table to Boolean.

**Database:**

**We first created an EC2 instance in the AWS and installed postgres.**

**We then downloaded pgAdmin4 and connected it to that server to easily interact with it using a good User Interface.**

Graphical user interface, application

Description automatically generated

As we can see that we have created 10 Tables based on our Class Diagram in part 2 and added the data in it. We have also added clustering and indexing in the Table. We will be talking about the schema of all the Tables as we have added in the pgAdmin.   
  
Download postgres using command shell in EC2 instance.  
Download pgadmin and create a server and connect it to postgres by giving the IP Address.

**CUSTOMER:**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

**EMPLOYEE:**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

**HISTORY AND VEHICLE:**

Graphical user interface, application

Description automatically generated

**Table

Description automatically generated**

**Table

Description automatically generated**

**CLAIMS:**

Graphical user interface, text, application

Description automatically generated

**Graphical user interface, application

Description automatically generated**

**INVOICE:**

**Graphical user interface, application

Description automatically generated**

**FINANCIAL INST:**

**Graphical user interface, application

Description automatically generated**

**POLICY:**

**Graphical user interface, application

Description automatically generated**

**PRODUCT:**

**Graphical user interface, application

Description automatically generated**

**TRANSACTION:**

**Graphical user interface, application

Description automatically generated**

**OPTIMISATIONS:**

1. **We have replicated the data for fault tolerance. We have 3 instances of the DB running.**
2. **We have used clustering in the database also for:  
   Scalability, availability, and future monitoring.**
3. **We have also added 3 indexes:  
   We have a State\_Index in Customers table so that Customers from a particular state can easily be accessed efficiently.  
     
   Graphical user interface, application

   Description automatically generated  
     
   We have an SSN Index in Employee as this can be used to segregate Employees based on certain area codes. This is a clustered Index as we want the physical order in which table records are saved to be SSN based.  
     
   Graphical user interface, text, application

   Description automatically generated**

**We also have a Start\_date\_Index in the Policy Table so that we can efficiently access policies based on ascending date order.**

**Graphical user interface, application

Description automatically generated**

**We have included the DB.txt file which contains the entire postgres data dump.**

**Machine Learning Model Creation:**

**Business Use Case:**

We have done Cardiovascular Disease Prediction based on the Age, habits, Cholesterol, etc. of the Customers. Our Model gives a probability of a person having Cardiovascular Issues. Based on this probability, we determine the Car Insurance rate as the [person who is more prone to such disease are prone to utilize the insurance and hence the premium may be high.  
  
**Big data Ideas for long term benefits:**

We have downloaded the dataset from the S3 bucket where we uploaded it in part 2. We also plan to upload the model created to the cloud so that it can be used over the internet.

Furthermore, the model will keep on training based on the enormous amount of new data that is coming which would increase the accuracy of the model.

We can use services like Kafka for the stream of the data to make a queue and pass the data to the model when it is ready and at the same time signals the Customers that we have received the data.

**Data Pre-processing:**

First, we download the dataset as csv and load it as a data frame. We then see the basic properties of the dataset like the shape and description.   
  
We then remove the NULL values from the dataset.  
We replace the NULL values with the mean values of the respective columns.

We then check for unique values for each column.

**Feature Engineering:**

We did an exploratory Data Analysis to see the features that are highly correlated and removed them to improve the accuracy of the model.

Chart

Description automatically generated

**We Used SMOTE to balance the data**

**Train Test Split:**

We have divided the dataset of 80% and 20% training and testing data.

**We then created a logistic regression model whose accuracy was 83.2%. This Logistic Regression gives us the percentage probability of the likelihood of the disease based on which we decide the premium prices using predictPremium function.**

**We have attached the dataset, the model pkl file which we have trained, the jupyter notebook, the dataset, along with the figures in the pdf file.**