Karla Dominguez Melo

Stanislav Mykhaylichenko

Xingyi Li

Japman Kaur Dhaliwal

CISC5950 Big Data Programming

05/08/2020

**Project 2 Part 2**

This project was developed and tested using a local cluster consisting of 3-nodes with Hadoop Distributed File System, running Ubuntu 16.04 LTS in GoogleCloud.

In this project, we are going to use a spark cluster for data processing jobs. The project consists of 4 parts.

For Part 2, we will predict whether the patient will have heart disease in ten years based on 15 attributes using the Framingham Heart Dataset.

The data contains anonymous information such as age, sex, education, heart rate, etc. The goal is to train a binary classifier to predict the possible heart disease in 10 years. There are 4240 instances and 15 attributes in the dataset. The data contains a good blend of categorical, numerical and missing values.

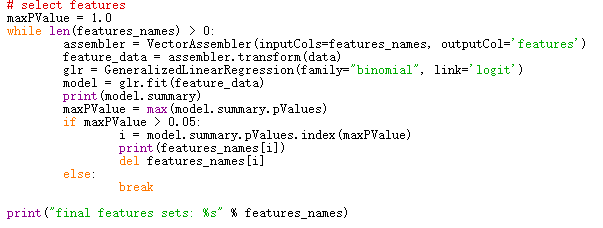
In this part, we were asked to use the logistic regression algorithm on Spark ML/MLlib to train the model and evaluate it.

**Data Preparation**

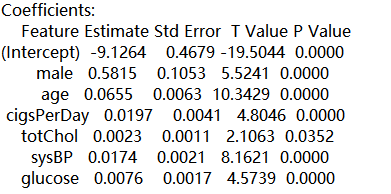
The education column is discrete value, and not related to the heart disease prediction, so we decided to drop this column. There are about 400 rows containing the missing values. Since it is only 12 percent of the entire dataset, we will drop those rows.

**Feature Selection**

To select the significant features, we will use logistic regression and the p-values of each feature. PySpark ML library has GeneralizedLinearRegression model and its summary contains the P values. We started with all features, and train GeneralizedLinearRegression model with family binomial and link logit on all data. Then we check p-values of all features, and remove the feature with the largest p-value if p-value is above 0.05. After feature selection, we should only select features with p-value below 0.05.



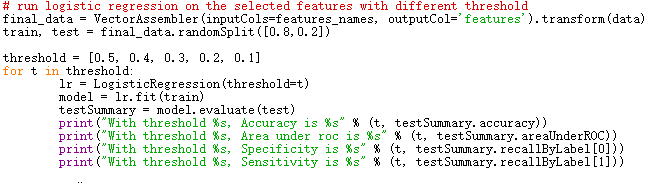
The diagram below shows the final feature set we selected.



**Analysis**

To train the model, we select the data with only selected features, and split the dataset into training data(80%) and test data(20%).

After both data sets were prepared, we used the LogisticRegression in PySpark ML lib to train model with training data, and evaluate the model on the test dataset.



**Result**

We explore the model performance with different threshold, and here is the result.

