**Module 2 - Fundamentals of AI**

**Course Number:** EAI 6000

**Academic Term:** Fall 2020 CPS Analytics  
**Instructor’s Name:** Kasun Samarasinghe

**Assignment Completion Date:** 11-08-2020

**Submitted By:**

Dhiren Vasudev Pagrani

Sunil Raj Thota

Shivani Sharma



**INTRODUCTION**

Detecting disease associated genes and organizing the disease type using gene expression data is very useful area in machine learning and artificial intelligence. Several models such as logistic regression model and support vector machines (SVM’s) have used in this area.

In this report, we will apply Linear and Logistic Regression to analyze the variability between values and predict results in a more accurate way.

Linear regression is a continuous model used when we want to predict the value of a variable based on the value of another variable. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. The linear regression model can be represented by the following equation.



* *Y* is the predicted value
* *θ*₀ is the bias term.
* *θ*₁,…,*θ*ₙ are the model parameters
* *x*₁, *x*₂,…,*x*ₙ are the feature values.

The above hypothesis can also be represented by:

Image for post

* *θ* is the model’s parameter vector including the bias term *θ*₀
* *x* is the feature vector with *x*₀ =1

Logistic is an appropriate regression analysis to perform when the dependent variable is dichotomous or binary. This is often a predictive analysis and it’s used to define data to describe the connection between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. It is primarily accustomed to make classifications into various categories like 1/ 0, Yes/ No, True/ False or Failure/ Success, etc. It’s a classification algorithm. In simple words, it predicts the probability of occurrence of an incident by fitting data to a logit function.

p =

It can handle any number of numerical and/or categorical variables,

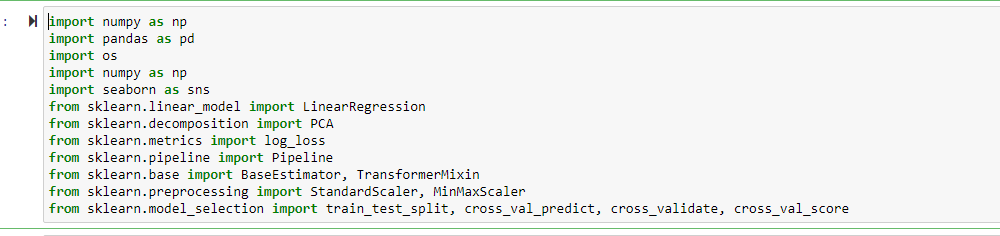
p =

Choosing the right predictor variables and avoiding the use of highly correlated variables are the main challenges to deal with. One has to consider the Risk Analysis Assumptions on the relationship between input and output variables as well to handle the analysis well. Few examples where logistic regression is applied are Fraud Detection, Credit Scoring, Email Spam, Real Estate Prices, Hotel Booking, Object Detection and etc.

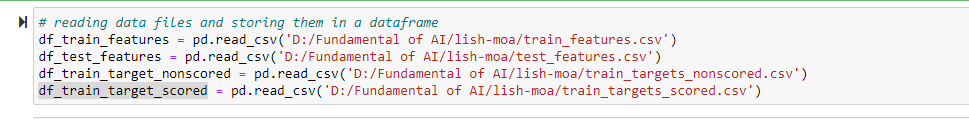
In this project, we will be assessing the linear and logistic regressions algorithms.

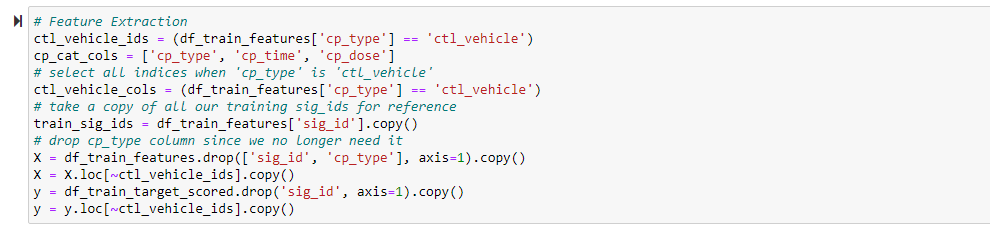
**Analysis**

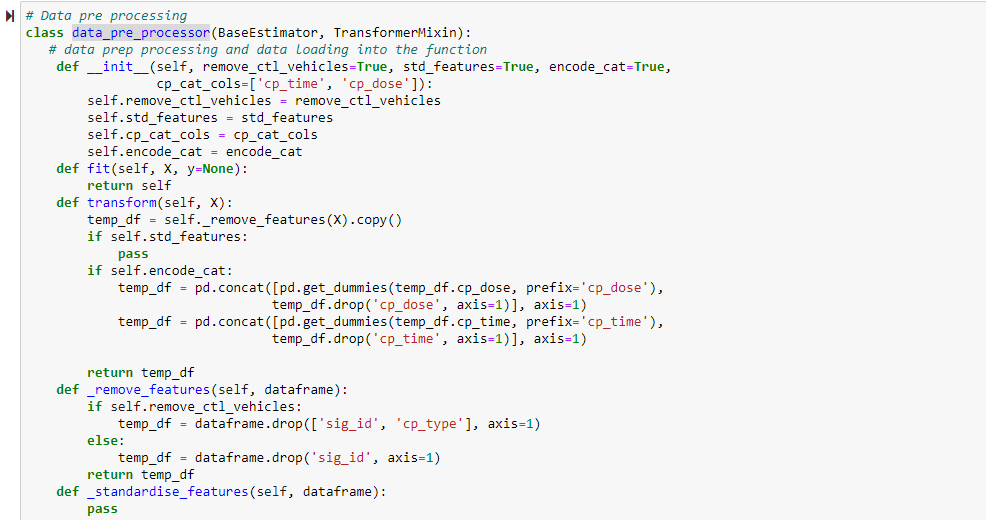
Here is the analysis for the linear model. Following are the libraries we need for linear model analysis on our selected dataset:

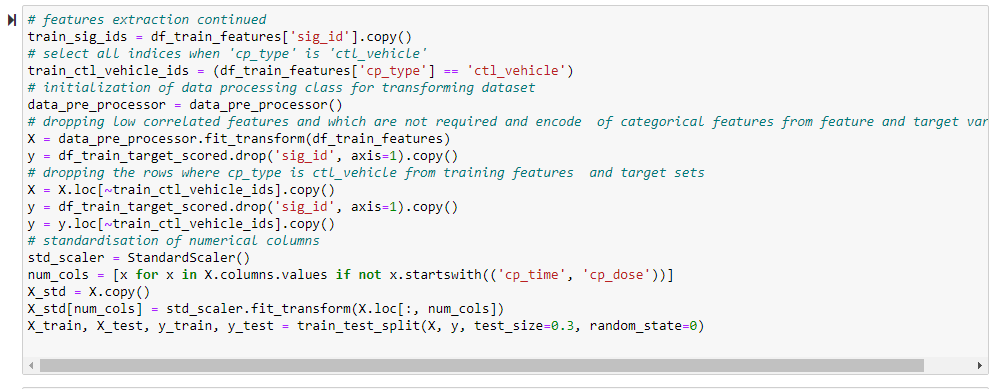


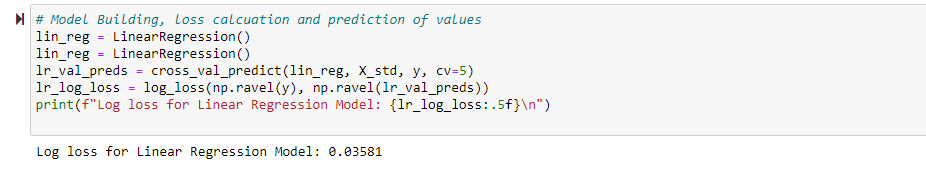
Now following is the logic for storing the dataset in dataframes:

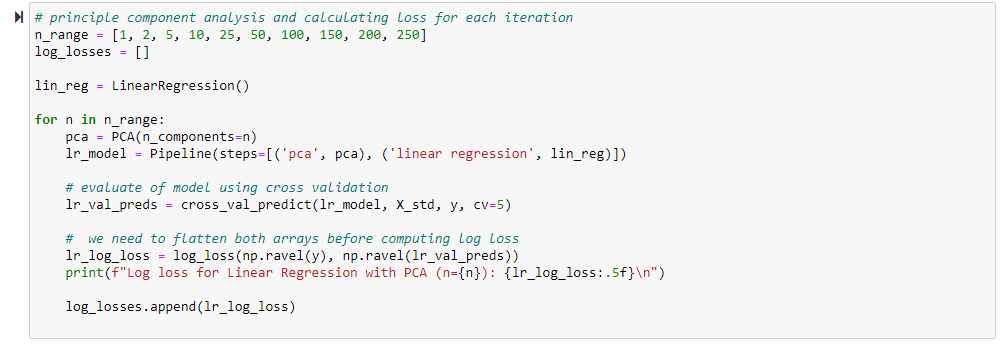


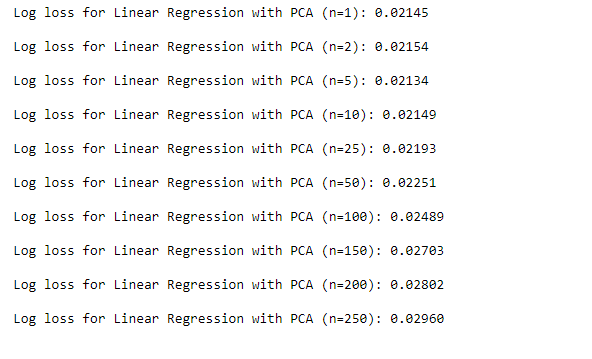




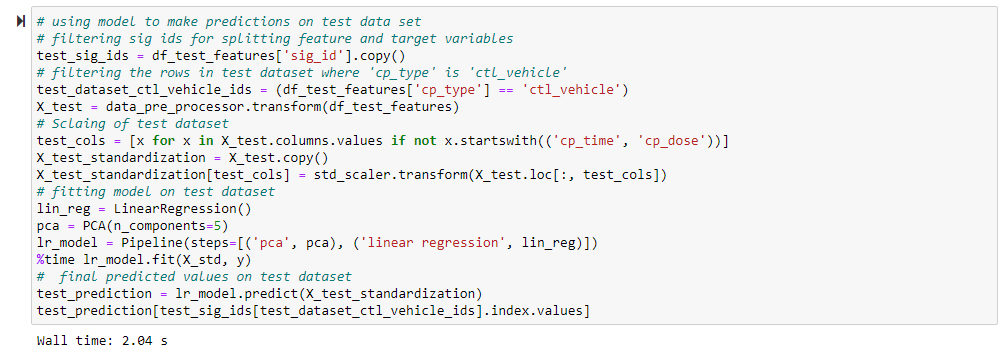


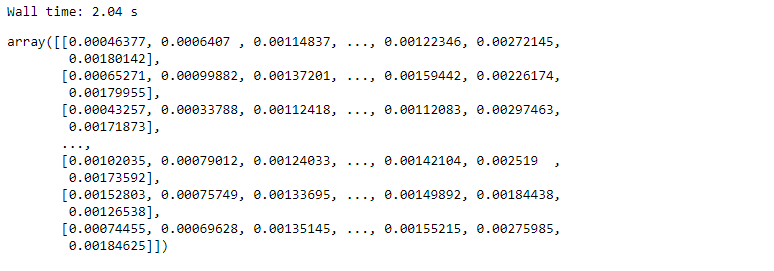






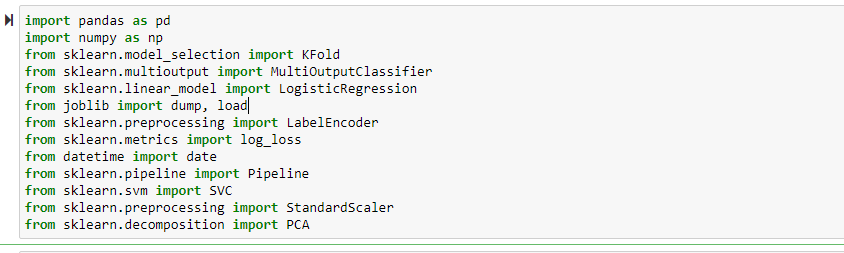




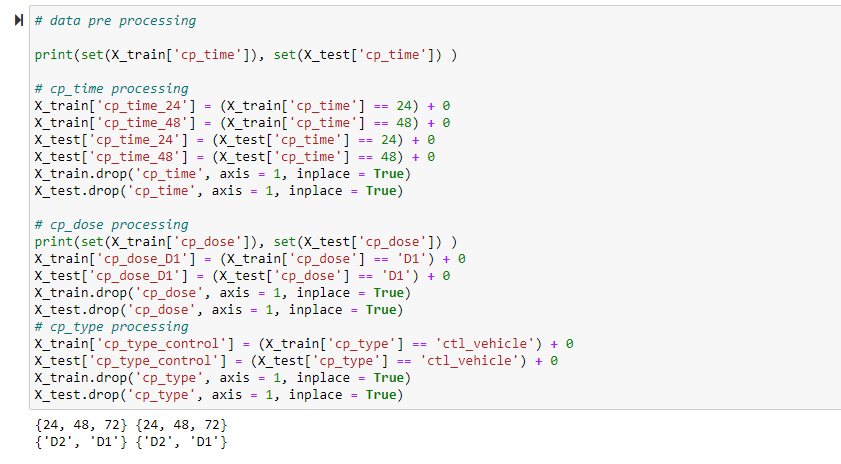




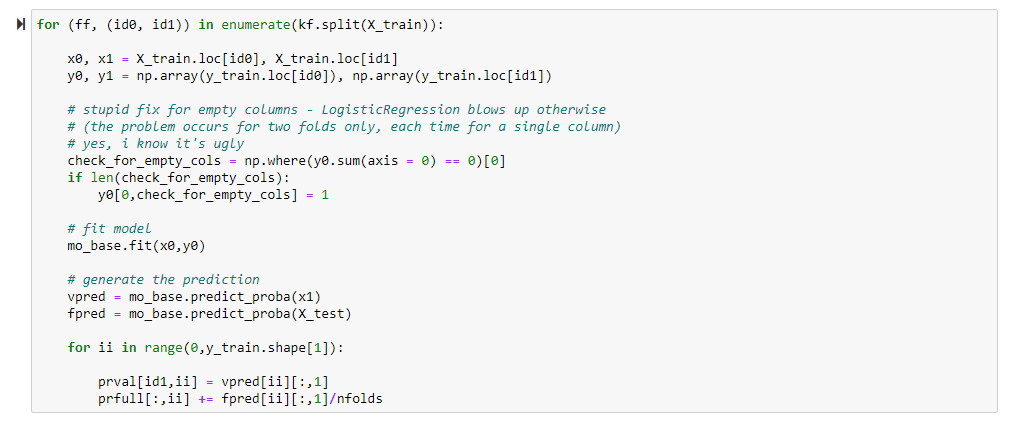
Now coming to logistic model logic:

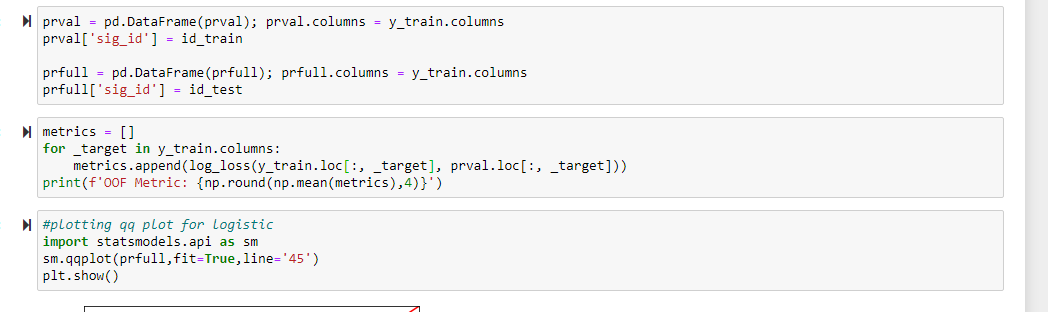






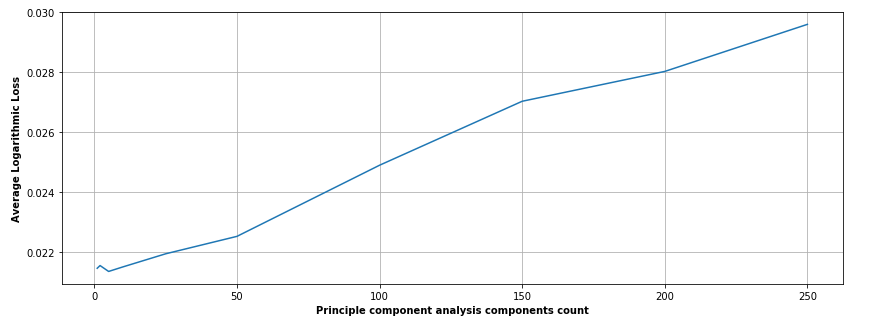






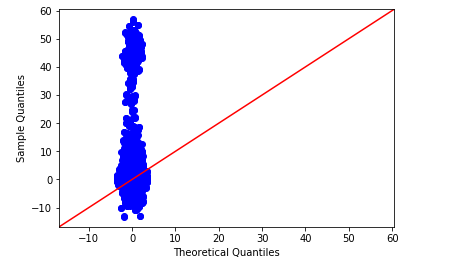
**Conclusion**

Following are the visualizations created for the linear and logistic regression models.

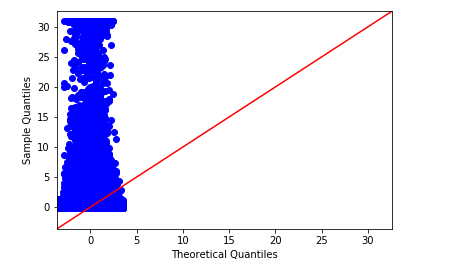


From the above curve we can see that there is little non linearity in the data set and hence we build logistic on top of this model in order to reduce it.

Below is the qq plot representation for the linear model.



QQ plot for logistic regression model is as follows:



**References**

1. Real Python. “Linear Regression in Python.” Real Python, Real Python, 7 Nov. 2020, realpython.com/linear-regression-in-python/.
2. Animesh Agarwal, (2018), Linear Refression using Python, Retrieved from <https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2>
3. Jaadi, Z. (n.d.). A Step by Step Explanation of Principal Component Analysis. Retrieved November 09, 2020, from <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>
4. Varshney, P. (2020, October 17). Q-Q Plots Explained. Retrieved November 09, 2020, from <https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0>
5. Interpretation of the Fitted Logistic Regression Model. (2005). Applied Logistic Regression, 47-90. doi:10.1002/0471722146.ch3
6. Model-Building Strategies and Methods for Logistic Regression. (2005). Applied Logistic Regression, 91-142. doi:10.1002/0471722146.ch4