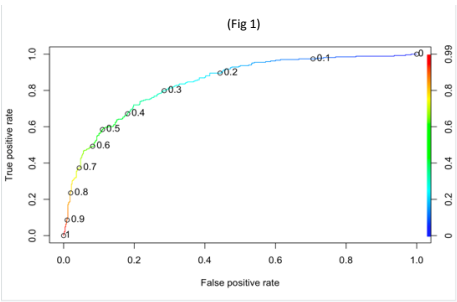
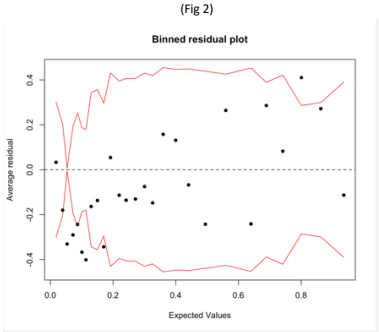
REGRESSION ANALYSIS BASED ON THE ORIGINAL DATASET:

Firstly, regression line has been fitted to the original dataset. Checking on the deviance residuals, it was evident that most of the diabetes prediction from Pima Indians Diabetes Dataset has neither been over estimated nor underestimated by looking at the median of the residuals. A higher null deviance of "993.48" on 767 degrees of freedom says that there is a need of more than a single variable to predict the diabetes data and only intercept is not enough. In addition, on considering 8 variables already predicted in the model, the residual deviance is comparatively lower than the null deviance. But not perfectly alright for the model to be a perfect fit for the Pima Indians Diabetes Dataset as a higher residual deviance means that there is a higher variation among the estimated data than predicted.[1] A greater AIC that has been observed in this case tells that this is not a better fit for the Diabetes dataset. Further, taking a 5% risk in determining the significance level, the variables "pregnant", "glucose", and "mass" are statistically significant during the process of estimating the presence of diabetes in an individual. Consequently, taking the log odds of probabilities into consideration, being pregnant or old, having high glucose concentration, BMI, triceps skin thickness, is more likely to prone to diabetes in Pima Indians while blood pressure and insulin is less likely.
Secondly, an AUC of 0.83 using the ROC curve tells that this is not the perfect model but can be considered. A threshold of 0.2 has been taken as part of research of the Pima Indian Diabetes Dataset. [Fig 1] The reason for this will be that considering this diabetes dataset minimization of the prediction of a negative test report being a positive in this dataset as diagnosis of diabetes is necessary from the early stages. The aim is to always have a maximized true positive rate against false positive rate based on the data set. For this research, it is clearly shown from the figure 1 that the true positive rate is 0.90 while the false positive rate is 0.44 at an alpha value of 0.2.

(Fig 1)



(Fig 2)

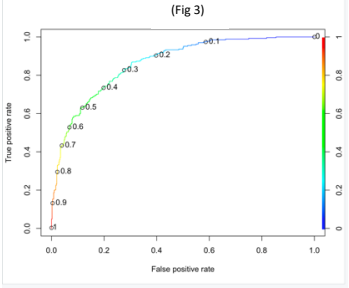


Thirdly, different residual plots can be plotted to check the residuals of the Dataset. But, as Pima Indians Diabetes dataset has a binary outcome and includes the logistic regression analysis, the QQ plots, residual vs fitted plots, etc. do not give proper results. To compensate that, binned residual plot analysis is used. Binned Plot gives the model performance based on the data, after the division into bins, lying in the red bands (95% Confidence Interval). [fig 2] These bands are called as the SE bands. Here, we see there are few bin categories outside the bands on the left side giving us the result that this model doesn't look that reasonable. [2]
Lastly, for this model, a confusion matrix is fit to assess the errors with the prediction of the presence of diabetes in the individuals. Only 56% (sensitivity) of the data from the dataset have been classified correctly as positive while 89% as negative. This model has an accuracy rate of 67%. Further, the data collected is 39% (Kappa Value) reliable to all the attributes quantified in the Pima Indians diabetes Dataset. [3]

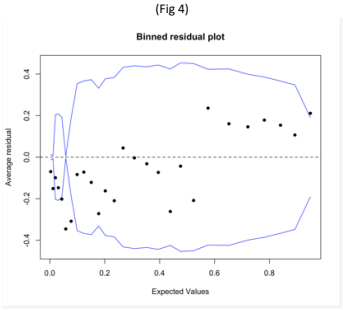
FITTED TRANSFORMED MODEL:

No missing data has been investigated for this dataset. Firstly, A model with first order interactions with all the other characteristics from the Pima Indians Diabetes Dataset gives a similar interpretation of deviance residuals to the model without any interactions. On the other hand, the Residual deviance is relatively lower, in accordance with the reduction of degrees of freedom, than the original model. An AIC of 735 is comparatively lower than that of the former model. The accuracy for the prediction of diabetes presence is 70% with sensitivity as 0.6 and specificity as 0.9. From the binned plot, we can see that there are not much bin categories than other fitted models outside the bands which might be a better fit for the dataset. [fig 4] Secondly, as the previous model includes a lot of interactions, looks tedious, and have a greater number of variables estimated, I would like to stick on to the variables with interactions that are statistically significant from the above-mentioned model. It is observed that for the latter model, the AIC is much lower than both models. But a higher residual deviance has been seen than the model with all interactions. A negative relationship has been observed between the diabetes and pressure, age, and triceps. An associated positive relationship is seen from with the number of times being pregnant, glucose levels, body mass, pedigree, insulin, pedigree and triceps, pressure and age.

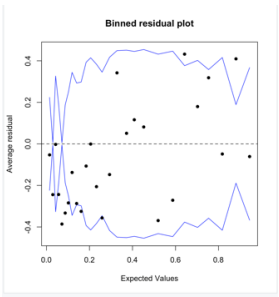
(Fig 3)



(Fig 4)



(Fig 5)



In addition, considering the AUC of 0.85 for the latter model being higher than the original model, says that this is a better model. According to ROC curve for the modified model, on choosing a threshold of 0.2 gives the false positive rate to be 0.4 and the true positive rate as 0.9. [fig 3] Checking the predicted quantities of the diabetes presence using the Confusion matrix gives us that the errors are comparatively minimized than the original model and the accuracy of the prediction has been increased by 2% but the prediction of true positives has been increased nearly by 4% leaving the specificity the same. Also, the reliability of the variables has become a little better with this model. Further, an analysis on residuals tells us that there are a few bin categories outside the bands that are almost same as the ones from the original model.[fig 5] Contradicting to the median from the deviance residuals, we can tell that most of the diabetes presence has been overestimated from the binned plot. Finally, looking at the likelihood ratios (that measures how good a model is) of each of the three models, the latter two models are statistically good compared to the original model. Further, the last model is better than the model with all first order interactions. [4]

CONCLUSION:

The model that is fit at the last is a better model than the original model considering the accuracy rate, the prediction errors, and the residuals. It is in addition concluded that a threshold of 0.2 has been chosen for better sensitivity and specificity. This came out to be better in the fitted model than compared to the original model. Further, Binned plots do not give much information on which is a better model between these two models.

References:

1. https://www.r-bloggers.com/2018/11/interpreting-generalized-linear-models/
2. https://rdrr.io/cran/arm/man/binnedplot.html
3. https://towardsdatascience.com/interpretation-of-kappa-values-2acd1ca7b18f
4. https://en.wikipedia.org/wiki/Likelihood-ratio_test