**A final paper explaining the problem, your approach and your findings in complete technical detail. Include ideas for further research, as well as up to 3 concrete recommendations for your client on how to use your findings.**

**Problem Statement**

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces leading to serious damage to many of the body's systems, especially the nerves and blood vessel.  The number of people with diabetes has risen from 108 million in 1980 to 422 million in 20141.

The global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 20141. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation1. While individual treatments is available for diabetes, insights from large chuck of data helps in prevention & identify people with high risk factors for Diabetes. At the same time diabetes can be prevented by lifestyle changes which will help in improving insulin sensitivity there by preventing Diabetes. Prediabetes means that your blood sugar level is higher than normal but not yet high enough to be classified as type 2 diabetes. Without intervention, prediabetes is likely to become type 2 diabetes in 10 years or less.2 Mayoclinic also states that certain lifestyle changes including eating healthy foods, getting more physical activity &loosing excess pounds can treat or even change they can reverse their condition. Hence to identify patients who are at risk of diabetes – this paper would like to explore and identify different markers of diabetes and identify the following

* Identify the markers of disease
* Create a model leverage early detection of Diabetes /increase predictability of diabetes
* Identify similarities in patients who have good sugar control as compared to other diabetic patients

**Approach**

The dataset had following. The breakdown of various data entities are as under:

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)^2)

7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1) where 0 means No Diabetes and 1 are Diabetic Group

Detailed Steps

1. In order to identify the key markers of disease the two groups were made –Diabetic & Non Diabetic Group and values were compared between the groups. Bar Charts along with Mean and 95% confidence variation was plotted between the two groups to show case the difference
2. Creation of Model using a. Logistic Regression (Diabetic & Non Diabetic Group) and Multinomial Logistic Regression using Diabetic, Non Diabetic Group.

Following Mathematical approach was used to create the model

**Logistic Regression**

…………………Eq 1

,

Taking Log of both sides

…………………….Eq 2

**Multinomial Logistic Regression**

Assuming there are three outcomes- Normal, Suspects & Diseased denoted by N,S & D then y3=log probability of suspects/Probability/(normal)

Then , and p(D)=………………….Eq.3

1. Selection of best Model using Sensitivity & Specificity Analysis

* True positive = correctly identified
* False positive = incorrectly identified
* True negative = correctly rejected
* False negative = incorrectly rejected

1. **Receiver Operating Characteristics Machine Learning**

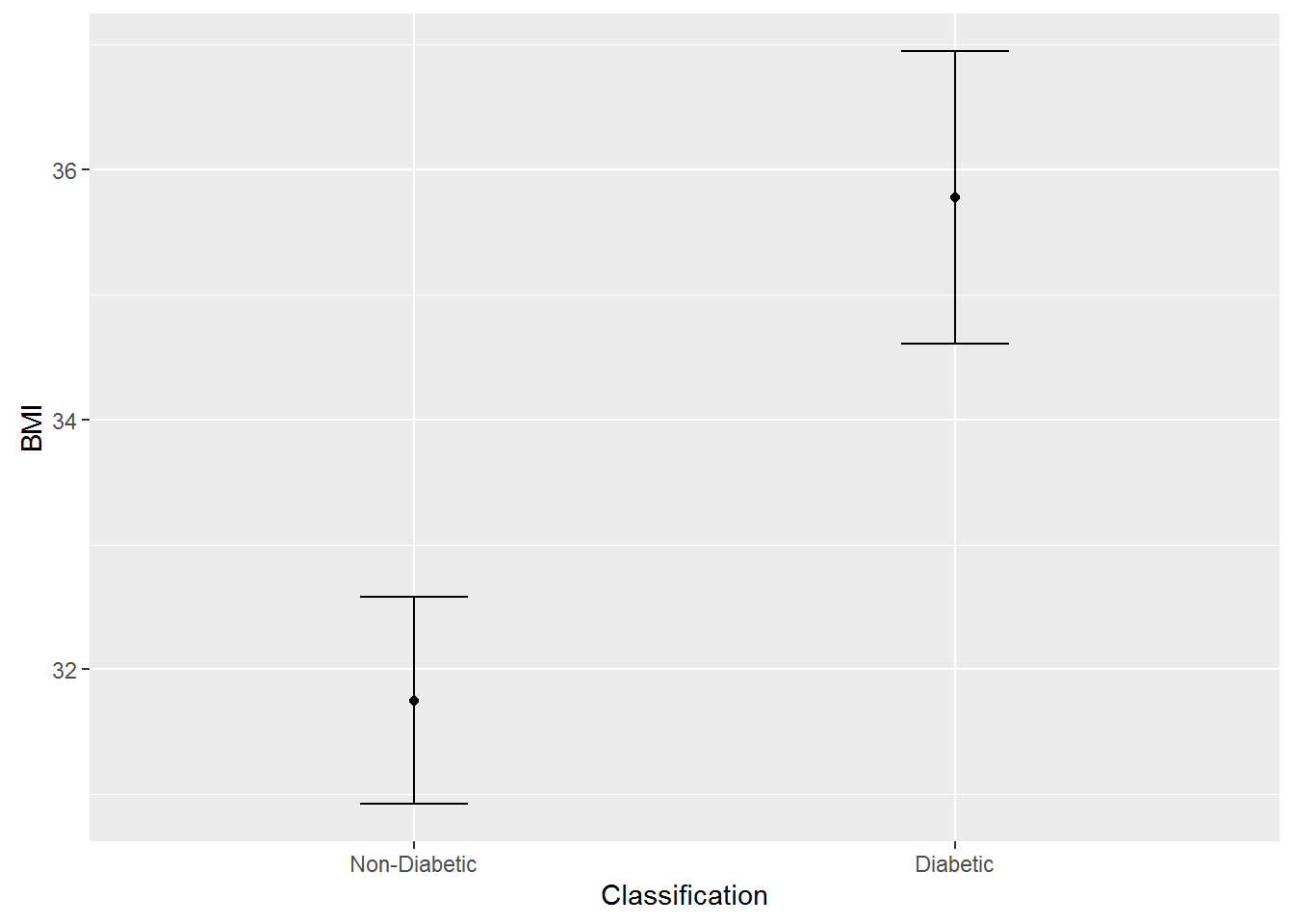
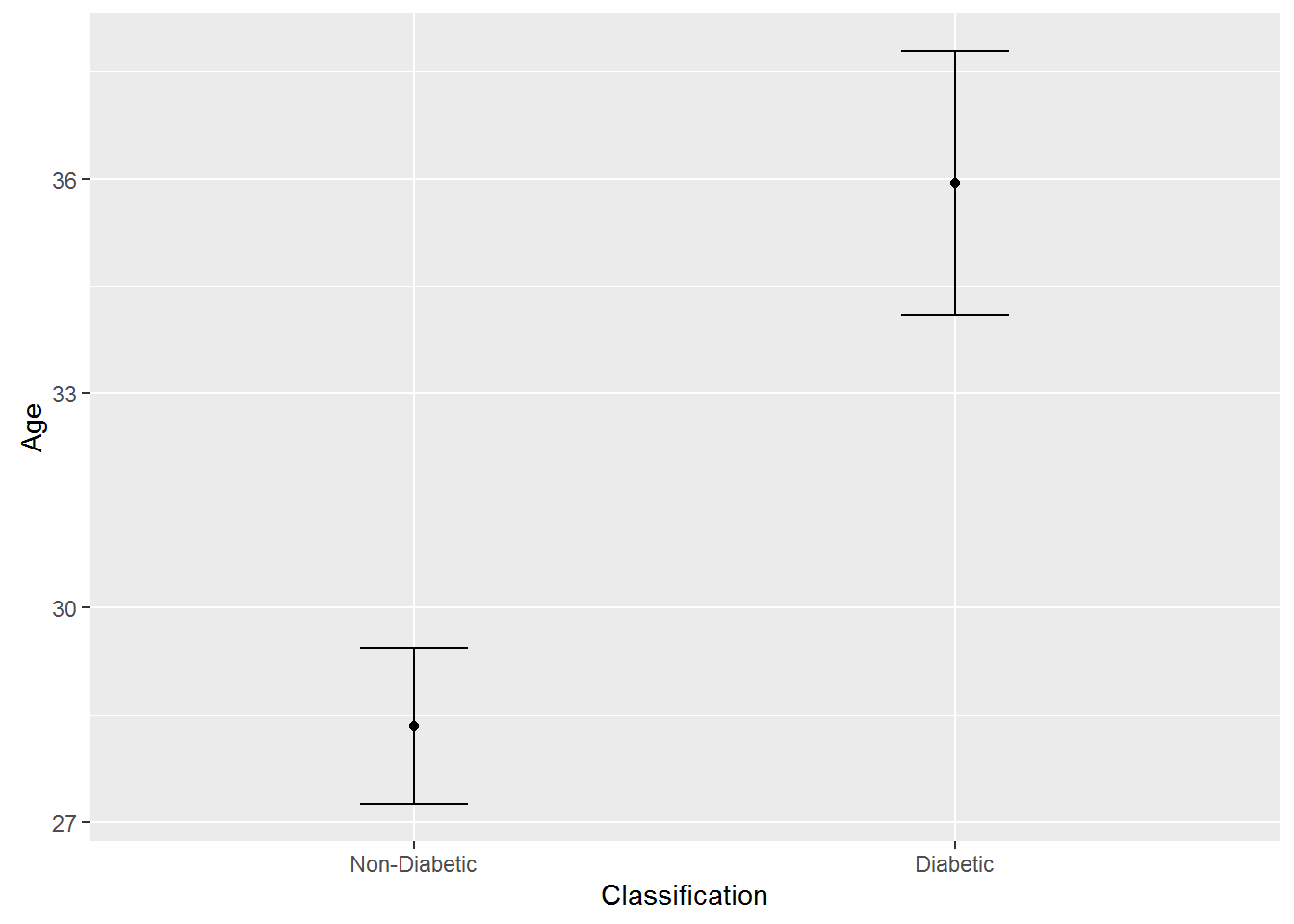
ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution.

…………………………Equation 4

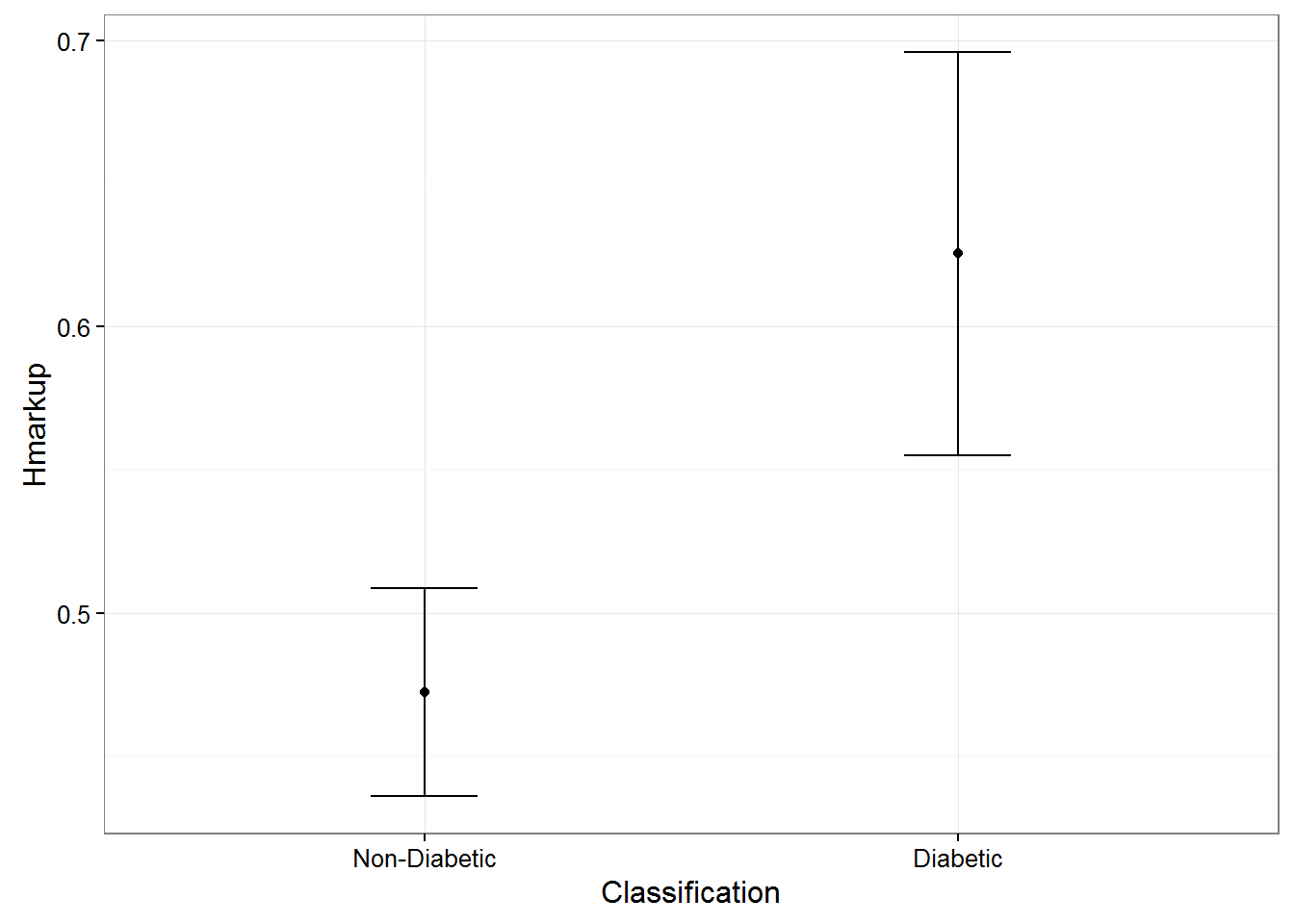
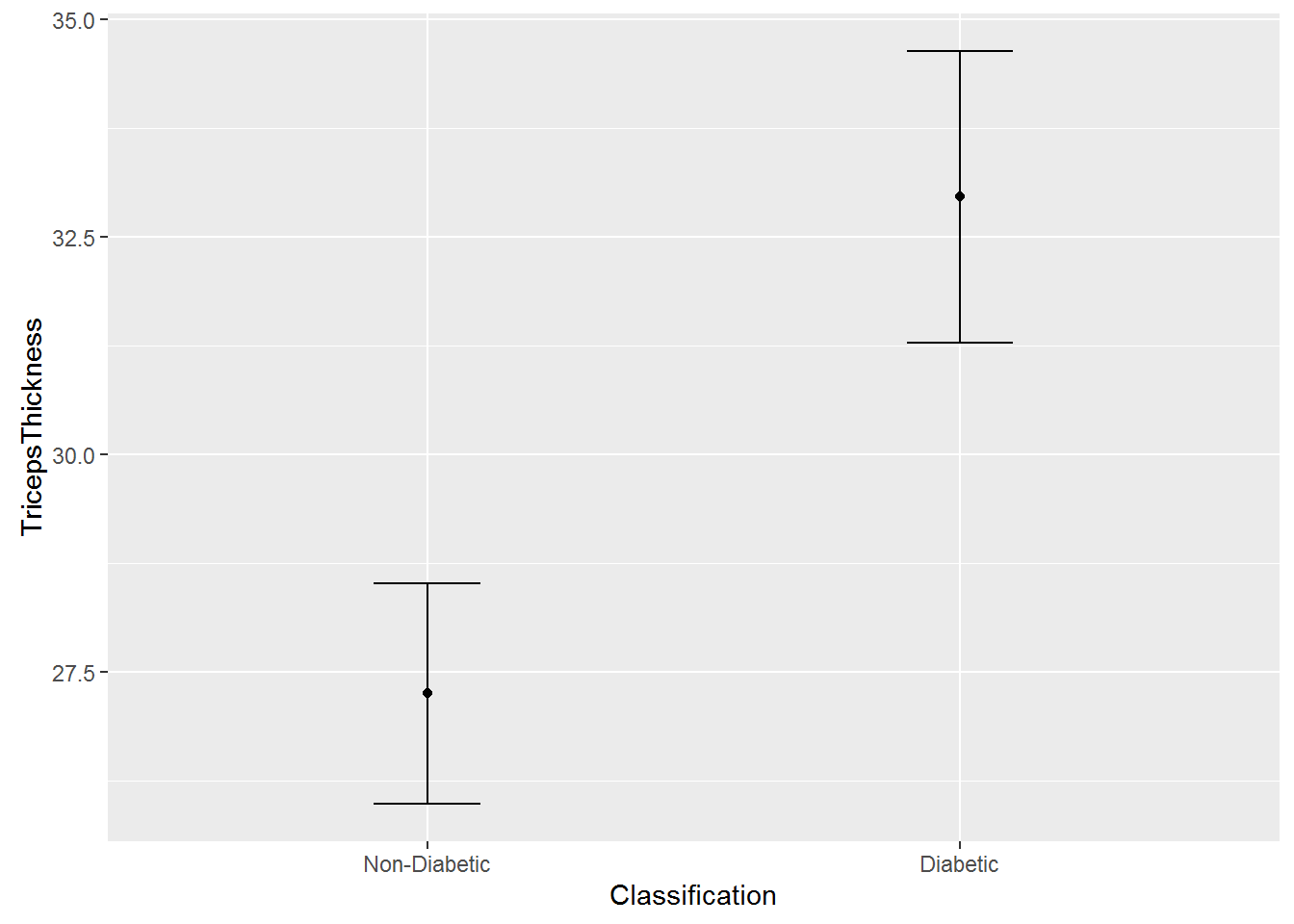
**Markedness =**

**Simulation**

It is evident from Figure 1,2,3,4 that subjects with diabetes had higher BMI,Hmark up,Tricpes Thickness and Age



**Figure 1- Mean Age & Variation Figure 2 Mean BMI & Comparison**

**Figure 3- HMarkup& Variation Figure 2 Mean Triceps Thickenss & Comparison**

Models & Machine Learning

Since there is a direct correlation between all of the above independent variable and binary variable various models have been designed. We would like to recommend top 2 models that have been derived from Training Dataset and which has been further validated in Test Dataset. ROC curves have been made to find and establish the efficacy of the model on the train data set.

Model1- All independent variables are continuous except BMI.

Two independent categorical variable has been created called- Overweight and Normal to make data driven decisions. Remaining variables are continuous variables so as to ensure we have higher sensitivity in Predicting the outcome variable. The outcome/confusion matrix based on t value 0.16 is as below

Insurance companies could use the model and predict the safest patients and hence reduced premium for patients with false results and Higher Premium for True Cases

At t value :0.1 Insurance Companies could play safe and assign the lowest premium for following candidates

|  |  |  |
| --- | --- | --- |
| Actual Outcome | Predicted No Diabetes /False | Predicted Diabetes /True |
| No Diabetes | 41 | 155 |
| Diabetes | 5 | 93 |

Model 2 –Independent variable are continuous.(BMI, Triceps thickness, Diastolic Pressure )

At t value: 0.15 Insurance Companies could play safe and assign the lowest premium for following candidates

|  |  |  |
| --- | --- | --- |
| Actual Outcome | Predicted No Diabetes /False | Predicted Diabetes /True |
| No Diabetes | 24 | 172 |
| Diabetes | 2 | 96 |

Model 3- Model 3 considers predicting diabetes by including Hereditary Markup, BMI, Triceps thickness & Diastolic Pressure.

At t=0.16 we have 100% sensitivity with average specificity. Hence at this point insurance company can ensure they are 100% safe in predicting diabetes without impacting customers.

This model does not consider interactions.

|  |  |  |
| --- | --- | --- |
| Actual Outcome | Predicted No Diabetes /False | Predicted Diabetes /True |
| No Diabetes | 42 | 154 |
| Diabetes | 0 | 98 |

Model 4( Considering Interactions among independent variables with Hereditary Markup, BMI, Triceps thickness ,age Diastolic Pressure. )

|  |  |  |
| --- | --- | --- |
| Actual Outcome | Predicted No Diabetes /False | Predicted Diabetes /True |
| No Diabetes | 89 | 169 |
| Diabetes | 1 | 97 |

Model 5 : Model considering independent variables Hereditary Markup, BMI,Triceps thickness & their interactions at t=0.13

|  |  |  |
| --- | --- | --- |
| Actual Outcome | Predicted No Diabetes /False | Predicted Diabetes /True |
| No Diabetes | 35 | 161 |
| Diabetes | 1 | 97 |

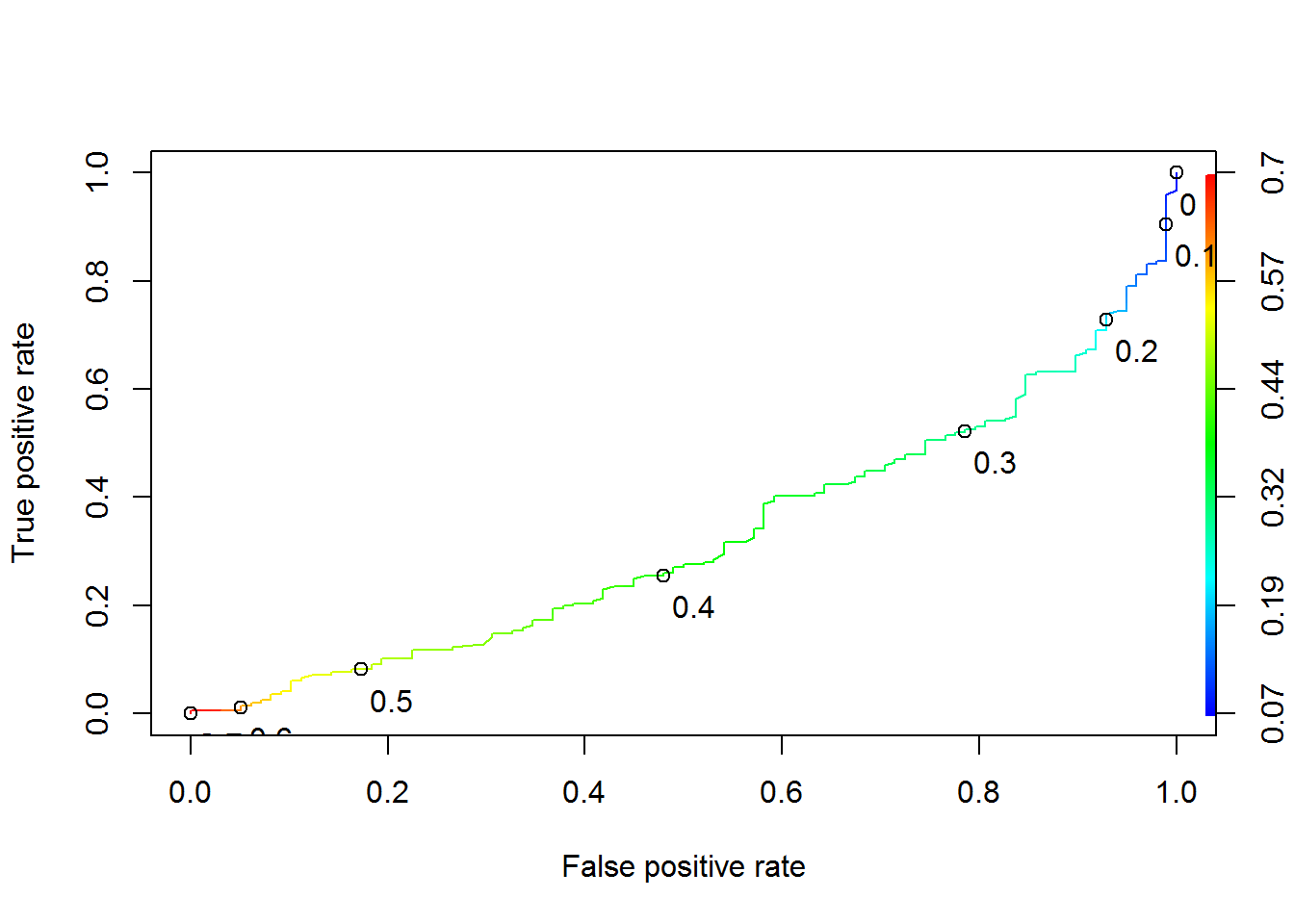
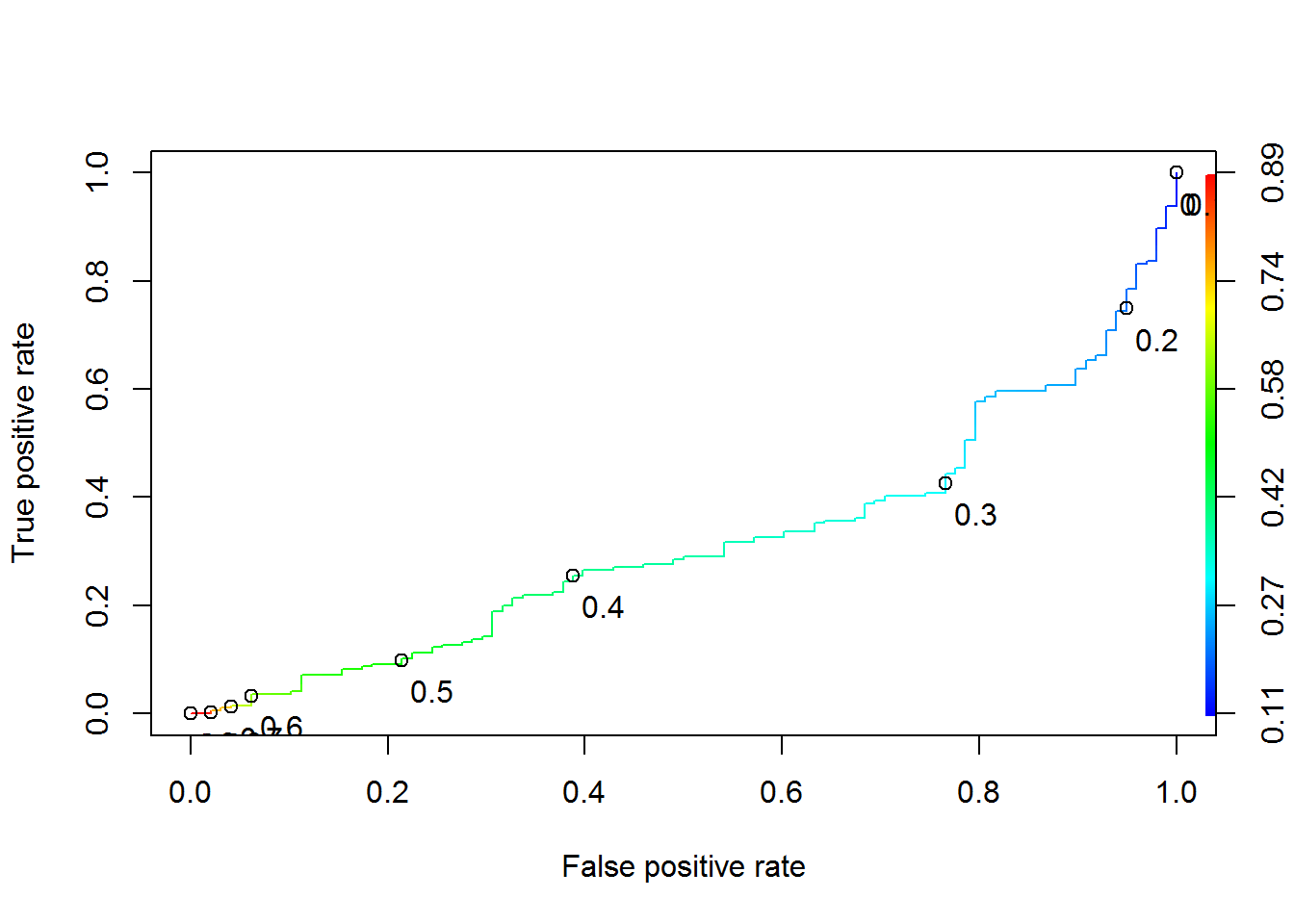
I

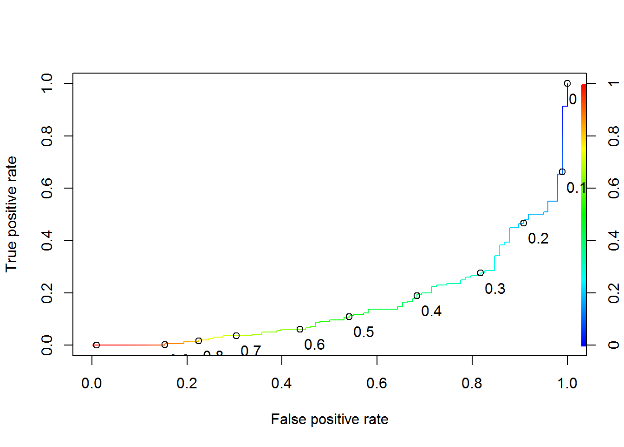
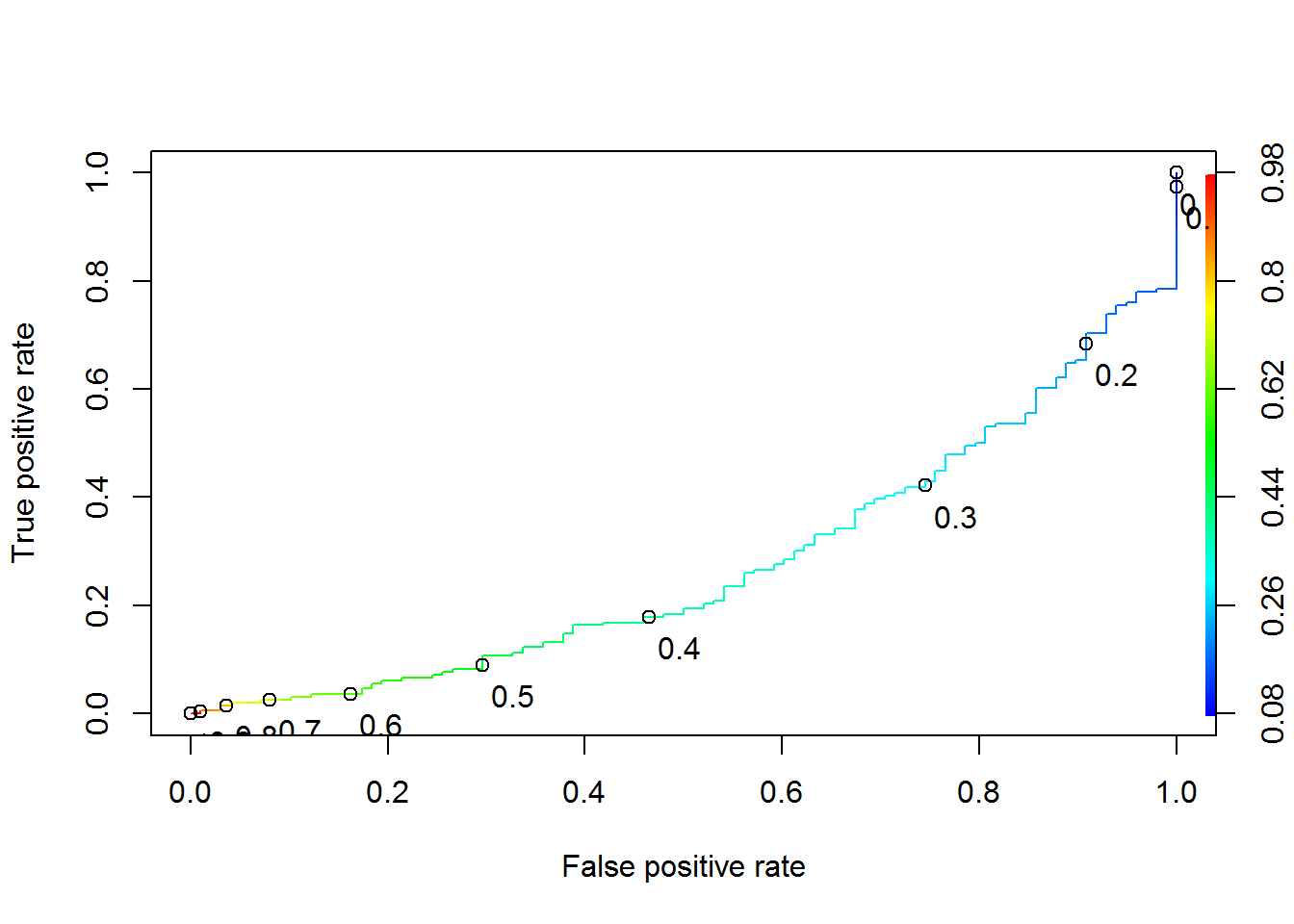
Model 6 –Considering BMI, Hmark up, Triceps Thickness ,Age, PP glucose Value into the model with t-0.4

|  |  |  |
| --- | --- | --- |
| Actual Outcome | Predicted No Diabetes /False | Predicted Diabetes /True |
| No Diabetes | 177 | 19 |
| Diabetes | 41 | 57 |

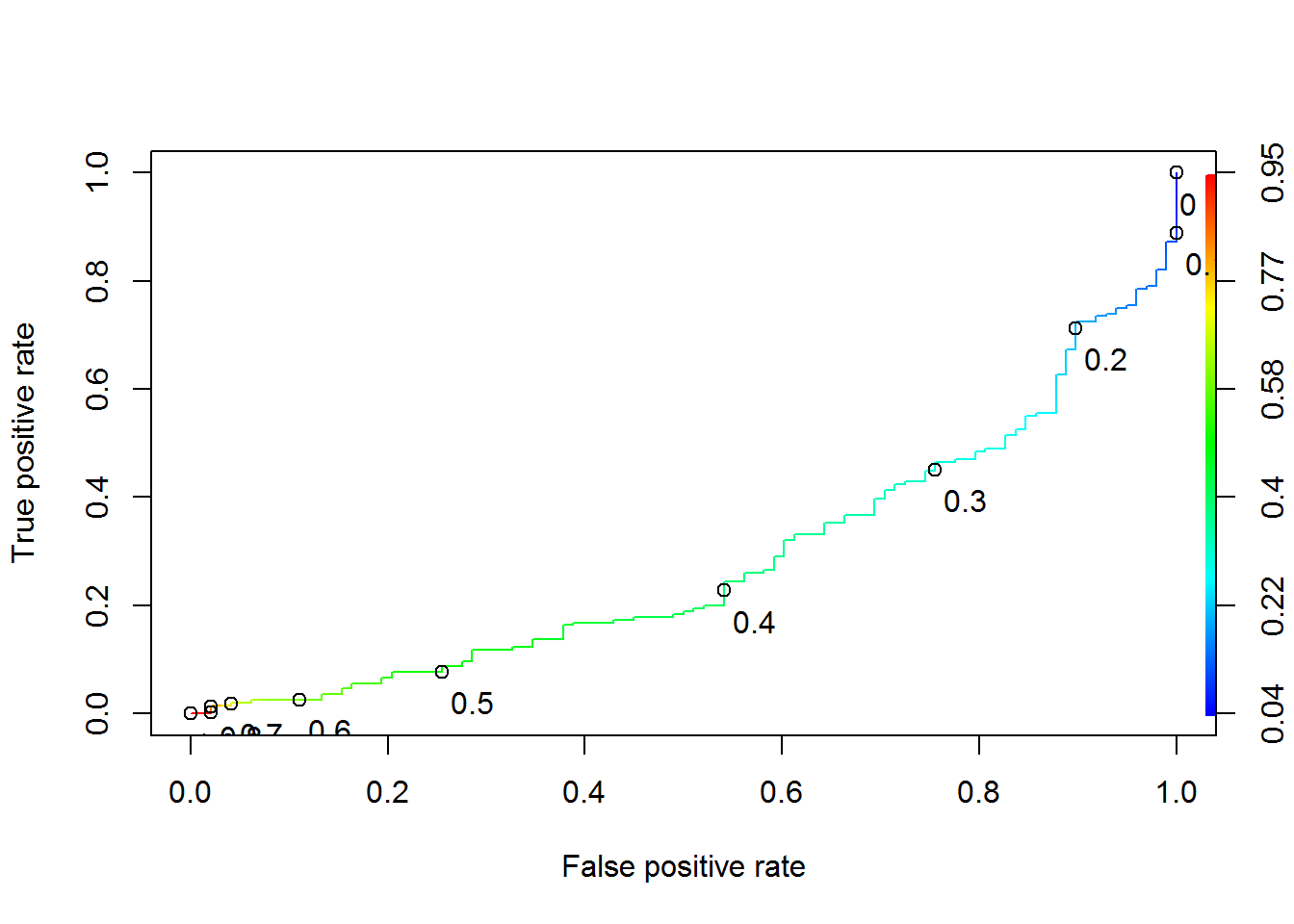
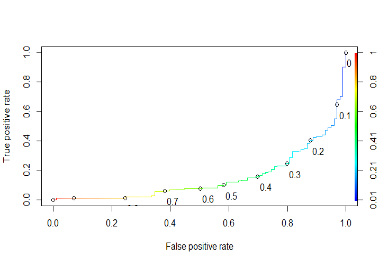
**ROC Curves for Different Models**

Model 1- Diastolic BP, BMI, Triceps Thickness Model 2 Diastolic BP, Factor BMI & Triceps Thickness

Model 3- Diastolic BP, BMI, Triceps Thickness & Hereditary Markup Model 4 DiastolicBP, TT, Age. HMark up & their interactions



Model 5-BMI, Triceps Thickness & Hereditary Markup and their interactions Model 6 (Considering Last know glucose LV) along with Model 3

Based on the ROC curves and Confusion Matrix Model 3 has highest Sensitivity with modest specificity. Hence firms could predict diabetes without their actual classification based on predictive model.

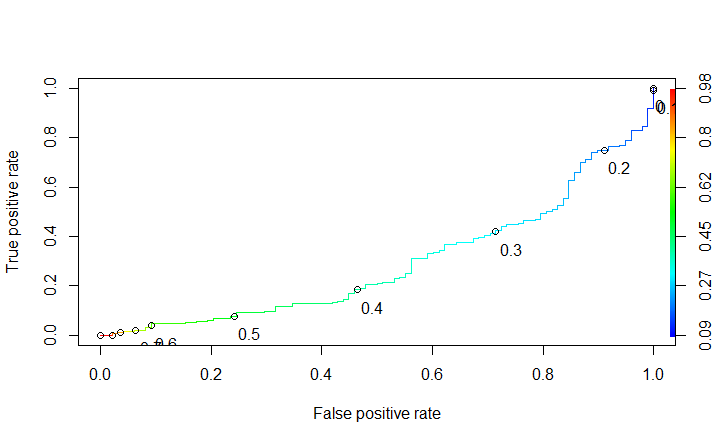
**Validation of Model in Training dataset for Model 3**

The model was run against test data to consider the efficacy of success of prediction. The model considered Diastolic BP, BMI & TricepsThickness. Hmark Up. Leveraging the model the prediction was able to predict 100% diabetes with about 48 cases being false Positive.

|  |  |  |
| --- | --- | --- |
| Actual Outcome | Predicted No Diabetes /False | Predicted Diabetes /True |
| No Diabetes | 18 | 48 |
| Diabetes | 0 | 32 |

Preciseness=== 40% and Markedness =39%

ROC curve to plot the sensitivity & specificity for different t values for Training Data



Model 3 : ROC curve for Test/Validation Data

Model 7: Model using Multinomial regression considering Pre-diabetes & Suspects.

All patients in data set whose PP glucose is between 140 & 200 and have not yet been classified as Diabetes have been considered as Prediabetic /Suspects .

Leveraged the following code to obtain the model

multinom(formula = out ~ BMI + Age + Hmarkup, data = pima.indian.diabetes10)

Coefficients:

(Intercept) BMI Age Hmarkup

Prediabetes -6.472219 0.06899522 0.06417325 0.8083701

Diabetes -7.349142 0.09914209 0.08947804 1.2518678

Hence =-6.42+.0689x+.06 x2+.808x3

Where x,x2 and x3 represent values for BMI,age &Hmark up respectively.

Creating another variable or Potential Suspects- Pre-Diabetes

|  |  |  |  |
| --- | --- | --- | --- |
| Actual | Predicted Normal | Predicted Pre-Diabetes | Predicted Diabetes |
| Normal | 198 | 0 | 29 |
| PreDiabetes | 21 | 0 | 14 |
| Diabetes | 29 | 0 | 65 |

Leveraging MultiLogit and Neural networks about 198/227 are predicted Normal among Normal Patients and 65/94 are predicted Diabetic in Diabetic Group.

Hence overall Model 3 and Multinomial Logistic Regression is able to predict with reasonable accuracy Diabetic Patients among the group.

The model based on Predictive learning algorithm can be leveraged by Insurance Companies to identify the members most likely to get Diabetes and determine the optimal message and channel of communication to connect these members with the appropriate premium and message to those systems. Predictive Analytics coupled with preventative interventions will improve the health of the highest-risk patients while reducing the cost.

With expenses rising and margins shrinking across the healthcare landscape, it’s critical that payers and providers alike leverage their medical data to gain predictive and informative insights into the health conditions faced by each member.

|  |  |
| --- | --- |
| Type of Client | How they could benefit from model |
| Insurance Companies | Increase Insurance cost for candidates who have higher predictability of Diabetes rationalizing revenue and prediction model for its earnings.  Improved target communication to the relevant Group |
| Health Companies/Hospitals | Better Cognition and advance detection of disease leading to better monitoring of patients |
| Patients | Better Preventive plan to Potential Suspects |

**Limitation**

The data in the dataset has only females and hence the marker for diabetes is useful only for females. The data in this set only has females of one origin. Hence multi facet analysis on ethnicity cannot be done. Certain data points have zero in BMI, Diastolic Pressure which is not biologically possible. These have been considered as outliers and have been removed from the analysis.

Another important factor which is not considered in the data set is amount of physical exercise performed. This is an important parameter which could help research substantiate on importance of exercise regimen on general health.

**Recommendation**

* 1. Insurance Companies could use the model 3 with t value 0.15 to identify patients who should have low insurance premium/least suspected patients.
  2. Hospitals and health Care companies who wants to target new Diabetic Medicine or new Diabetic Campaigns should ensure that campaigns should not target False Positives. Hence they should use Model 7 with t=0.4 to target definite patients
  3. Model 7 is recommended to be used by Healthcare Agencies to identify schemes/plan for Pre-Diabetics who are suspected to be Diabetics