Diabetes Prediction for HealthInsure Co.

Tim Hagan

2022-08-02

Contents

Summary

HealthInsure Co. is a small health insurance company looking to empower members to make informed decisions about their health by performing outreach for those who are at high risk of developing diabetes or having undiagnosed diabetes. A logistic regression model was built using the 2015 BRFSS Survey Data (6) to predict whether an individual was likely to have been diagnosed with diabetes or not. The performance of the model is strong as evidenced by the ROC curve, gains curve, and residual plots shown below. The provided prediction equation can be used to identify individuals who are at high risk of being diagnosed with diabetes and allow HealthInsure Co. to perform effective outreach.

Introduction

The CDC estimates that roughly 8.5 million adults in the US have undiagnosed diabetes (1). HealthInsure Co. wants to understand and perform outreach to members that it insures that are at high risk for having undiagnosed diabetes or developing diabetes. Uncontrolled diabetes can have serious complications for the individual including but not limited to eye complications, neuropathy, foot problems, nephropathy, and more (14). The ADA commissioned a study, *Economic Costs of Diabetes in the U.S. in 2017*, that estimates the total 2017 cost of diagnosed diabetes at $327 billion including $237 billion in direct medical costs and $90 billion in reduced productivity (2). The goal of the outreach campaign to members is to empower them understand their personal risk and provide resources so that they may choose to take action where appropriate. By performing creative outreach and providing resources for healthy lifestyle changes, HealthInsure Co. hopes to reduce the impact of uncontrolled diabetes on its members.

Goal: Using survey questions, build a model to predict the likelihood of an individual having diabetes.

Methods

To identify individuals that are likely to be diagnosed with diabetes, logistic regression was used since the target variable is binary (Yes/No) and the model is highly interpretable. While interpretability is not necessary for making strong predictions about which individuals are likely to have diabetes, insurance is regulated at the state and federal level and it may be necessary to show regulators how the model works to ensure the company is not in violation of any laws. Akaike information criterion (AIC) will be used to determine whether to include any variables that are *not* significant (more details below).

Target Variable: Diagnosed with Diabetes (1 if Yes; 0 if No)

Data

The Diabetes Health Indicators Dataset was downloaded from Kaggle in a CSV format (3). This data is a cleaned file generated from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 data using the code in this python notebook (4). All additional changes to the dataset were done in R. These involved the following:

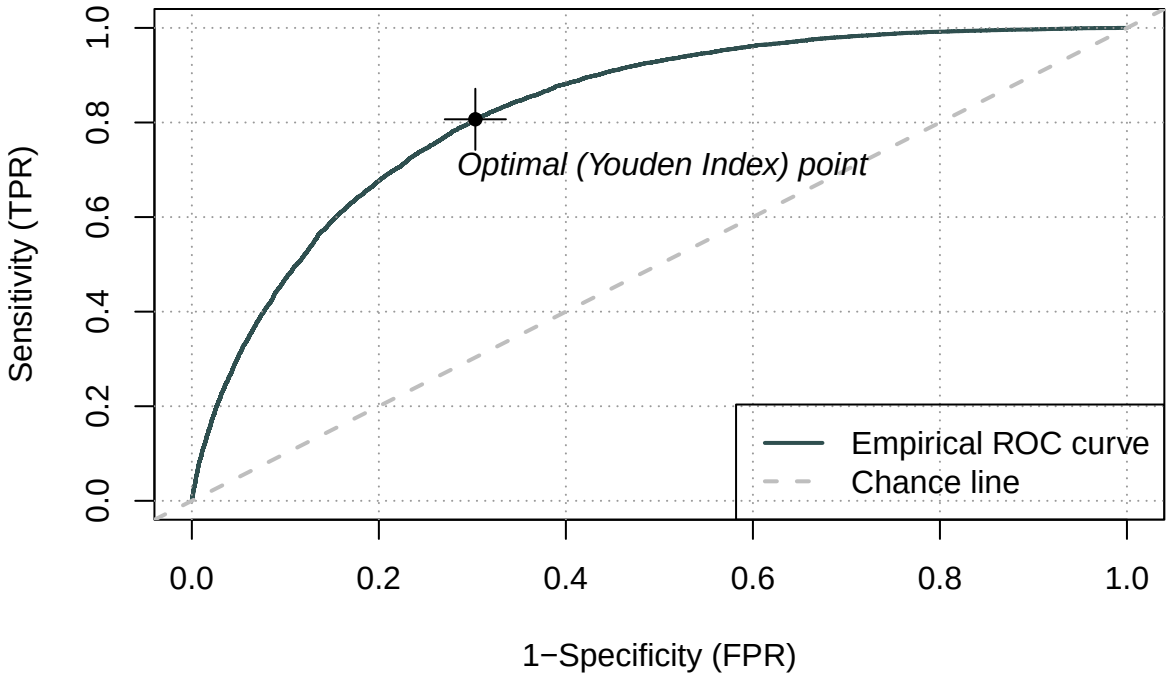
- Changing the target variable from (0/1/2; No Diabetes/Prediabetic/Diabetic) to (0/1; No Diabetes; Prediabetic or Diabetic)
- Recoding Diabetes_01 with levels 0/1 to Diabetes_YN with levels DiabetesNo/DiabetesYes
- Removing individuals without healthcare (AnyHealthCare == 0) from the dataset because HealthInsure Co. is by definition only doing outreach to individuals with health insurance
- Variable transformations for modeling purposes (more details in the Variable Definitions and Handling section)

We have discussed the goal of the model (predict whether an individual has diabetes), the kind of model used (logistic regression), and the data used to generate the model (BRFSS). Now we will turn to performance and examine it by looking at the ROC curve and gains curve on the test data set. Later in the model details section we will also look at marginal effect and residual plots by variable.

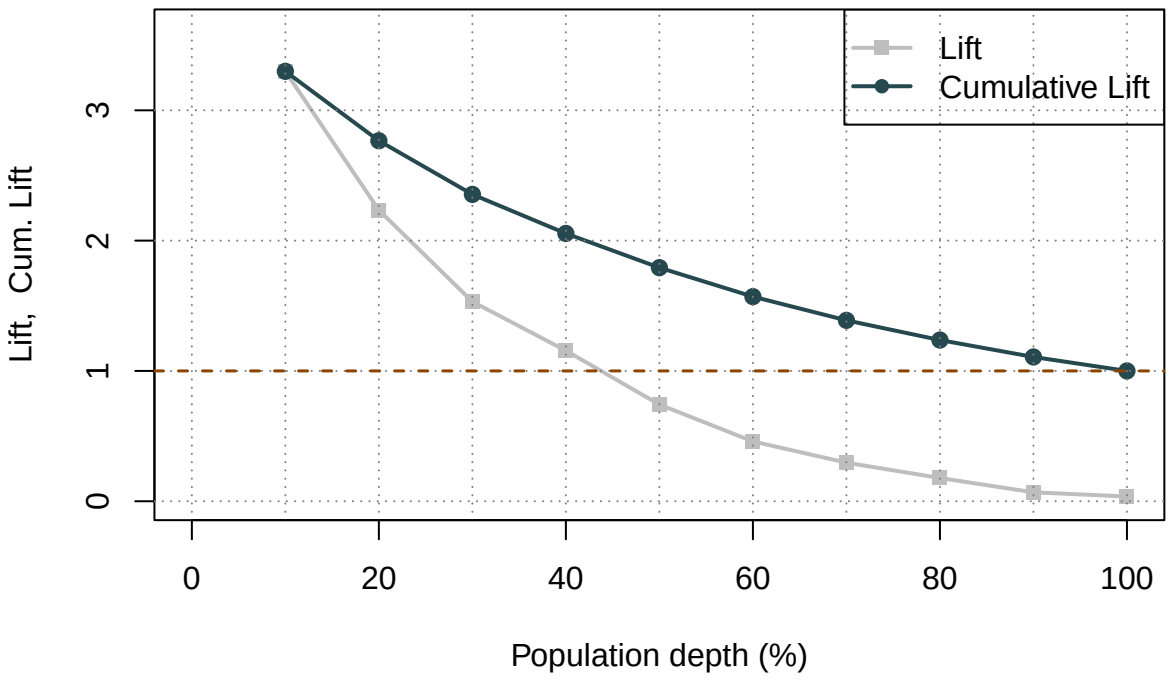
Performance

ROC curve The receiver operating characteristic curve in this case shows the ability of the model to identify cases of diabetes over chance as the cutoff threshold is varied. In choosing a cutoff, the trade off between true positive rate (TPR) and false positive rate (FPR) is shown by the curve with one measure of an optimal value shown (Youden Index). True positive rate (TPR) is equal to the number of cases

correctly identified cases of diabetes (True Positives) / the total number of cases of diabetes (Positives). The false positive rate (FPR) is equal to the number of respondents where the model predicted diabetes (False Positives) / the number of respondents without diabetes (Negatives).



Another way to look at performance is to look at gains table or curve. The below chart shows the lift and cumulative lift of reaching out to subsequent deciles of the population. For example, reaching out to the first decile on the test data (ordered highest risk of diabetes to lowest by model predictions) resulted in a lift of 3.3. This means that the number of diagnosed diabetes cases in that group was 3.3 times higher than expected (1/10th of the total population). For the second decile the lift was around 2.2 so in the first 2 deciles alone, more than 50% of the cases were identified.



Based on these results, the model performs well above chance and does a reasonably good job of ordering respondents from most likely to have diabetes to least likely. Now we will turn to looking at more details of the model such as variable treatment, model specification, marginal effects, and residual plots by variable.

Model Details

Variable Definitions and Handling

These definitions and a detailed handling can be found here (4) but will also be included below for the sake of completeness:

Diabetes
Diabetes_01 = (Ever told) you have diabetes or prediabetes?
1 if "Yes". 0 if "No" or "Only while pregnant".
This is the variable for which we are creating predictions.

High Blood Pressure
HighBP = Have you ever been told you have high blood pressure by a doctor, nurse, or other health professional?
1 if "Yes". 0 if "No".

High Cholesterol
HighChol = Have you ever been told by a doctor, nurse, or other health professional that your blood cholesterol is high?
1 if "Yes". 0 if "No".

Cholesterol Check

CholCheck = Cholesterol check within past five years?
1 if "Yes". 0 if "No".

Body Mass Index
BMI = weight (kg) / [height (m)]^2
For modeling, this variable was winsorized (capped at 50) to prevent overfitting at thin levels above 50. Additionally, natural splines (5) were used to capture the non-linear relationship between BMI and the target variable (Diabetes_01). An Excel document will be provided so that users may apply the same spline treatment to future data.

Smoking
Smoker = Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
1 if "Yes". 0 if "No".

Stroke
Stroke = (Ever told) you had a stroke?
1 if "Yes". 0 if "No".

Heart Disease
HeartDiseaseorAttack = Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
1 if "Yes". 0 if "No".

Physical Activity
PhysActivity = Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
1 if "Yes". 0 if "No".

Diet
Fruits = Consume Fruit 1 or more times per day.
Veggies = Consume Vegetables 1 or more times per day.
1 if "Yes". 0 if "No".

Alcohol Consumption
HvyAlcoholConsump = Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
1 if "Yes". 0 if "No".

Health Care
AnyHealthcare = Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service? 1 if "Yes". 0 if "No".
The data was filtered on this attribute. Because this project is for a health insurance company, only respondents that had health coverage were included in the study.

Avoidance of Doctor Visit Due to Cost
NoDocbcCost = Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?
1 if "Yes". 0 if "No".

General Health
GenHlth = How would you say that in general your health is?
Options: Poor, Fair, Good, Very Good, Excellent
The distance between each level is not apparent so this was treated as a factor with "Excellent" as the baseline.

Difficulty Walking or Climbing Stairs
DiffWalk = Do you have serious difficulty walking or climbing stairs?
1 if "Yes". 0 if "No".

Sex
Sex = Indicate sex of respondent.
1 if "Male". 0 if "Female".

Age
Age =
1 if 18 - 24
2 if 25 - 29
3 if 30 - 34
4 if 35 - 39
5 if 40 - 45
6 if 45 - 49
7 if 50 - 55
8 if 55 - 59
9 if 60 - 65
10 if 65 - 69
11 if 70 - 75
12 if 75 - 79
13 if 80 and up
Similarly to BMI, natural splines (5) were used to capture the non-linear relationship between Age and the target variable (Diabetes_01). An Excel document will be provided so that users may apply the same spline treatment to future data.

Education
Education = What is the highest grade or year of school you completed?
1 = Never attended school or only kindergarten
2 = Grades 1 through 8 (Elementary)
3 = Grades 9 through 11 (Some high school)
4 = Grades 12 or GED (High school graduate)
5 = College 1 to 3 years (Some college or technical school)
6 = College 4 years or more (College graduate)
This was treated as a factor because the difference in job types may be quite different from one level to the next.

Income
Income = What is your annual household income from all sources?
1 = Less than $15,000
2 = $15,000 to less than $25,000
3 = $25,000 to less than $35,000
4 = $35,000 to less than $50,000
5 = $50,000 or more
While the distance between levels is not constant, using this term in a linear fashion captured the effect of income well so it was treated as numeric. See Income residual plot below.

Prediction Equation

The model can be represented by the prediction equation below. Exponentiate both sides and divide the odds by (1 + odds) to find the predicted probability of having diabetes (Diabetes_01 = 1). Please note the excel sheet containing the mapping for the splines for BMI and Age.

$$
\begin{aligned}
\log\left[\frac{\widehat{P(\text{Diabetes\_01}=1)}}{1-\widehat{P(\text{Diabetes\_01}=1)}}\right] = & -7.41 \\
& + 0.63(\text{HighBP}) \\
& + 0.53(\text{HighChol}) \\
& + 1.18(\text{CholCheck}) \\
& + 0.13(\text{Stroke}) \\
& + 0.25(\text{HeartDiseaseorAttack}) \\
& - 0.03(\text{PhysActivity}) \\
& - 0.67(\text{HvyAlcoholConsump}) \\
& - 0.03(\text{Smoker}) \\
& - 0.02(\text{Fruits}) \\
& - 0.04(\text{Veggies}) \\
& + 0.04(\text{NoDocbcCost}) \\
& + 0.09(\text{DiffWalk}) \\
& + 0.23(\text{Sex}) \\
& + 1.66(\text{GenHlth}_{\text{Fair}}) \\
& + 1.25(\text{GenHlth}_{\text{Good}}) \\
& + 1.79(\text{GenHlth}_{\text{Poor}}) \\
& + 0.62(\text{GenHlth}_{\text{Very good}}) \\
& + 0.13(\text{Education}_2) \\
& + 0.01(\text{Education}_3) \\
& - 0.1(\text{Education}_4) \\
& - 0.05(\text{Education}_5) \\
& - 0.15(\text{Education}_6) \\
& - 0.07(\text{Income}) \\
& + 8.06(\text{naturalSpline}(\text{BMI\_win},\ \text{knots} = \text{c}(20,\ 40))_1) \\
& + 1.57(\text{naturalSpline}(\text{BMI\_win},\ \text{knots} = \text{c}(20,\ 40))_2) \\
& + 0.16(\text{naturalSpline}(\text{BMI\_win},\ \text{knots} = \text{c}(20,\ 40))_3) \\
& + 5.16(\text{naturalSpline}(\text{Age},\ \text{knots} = \text{c}(3,\ 5,\ 8,\ 10,\ 12))_1) \\
& + 0.31(\text{naturalSpline}(\text{Age},\ \text{knots} = \text{c}(3,\ 5,\ 8,\ 10,\ 12))_2) \\
& + 0.22(\text{naturalSpline}(\text{Age},\ \text{knots} = \text{c}(3,\ 5,\ 8,\ 10,\ 12))_3) \\
& + 1.19(\text{naturalSpline}(\text{Age},\ \text{knots} = \text{c}(3,\ 5,\ 8,\ 10,\ 12))_4) \\
& + 1.29(\text{naturalSpline}(\text{Age},\ \text{knots} = \text{c}(3,\ 5,\ 8,\ 10,\ 12))_5) \\
& + 1.9(\text{naturalSpline}(\text{Age},\ \text{knots} = \text{c}(3,\ 5,\ 8,\ 10,\ 12))_6)
\end{aligned} \tag{1}
$$

Model Specification

The only variables that were not statistically significant (alpha <= .05) were Education (each level), Fruits, NoDocbcCost, and Smoker. AIC (Akaike information criterion) assesses the "goodness of fit" of a model while penalizing additional complexity (i.e. variables) added to the model. The lower the AIC, the better the model. This was used to determine the "best model".

```
# AIC of the full model with all variables
AIC(full_model)
```

```
## [1] 132236.5
```

```
# AIC of the model without Education included
AIC(noEducation)
```

```
## [1] 132284.3
```

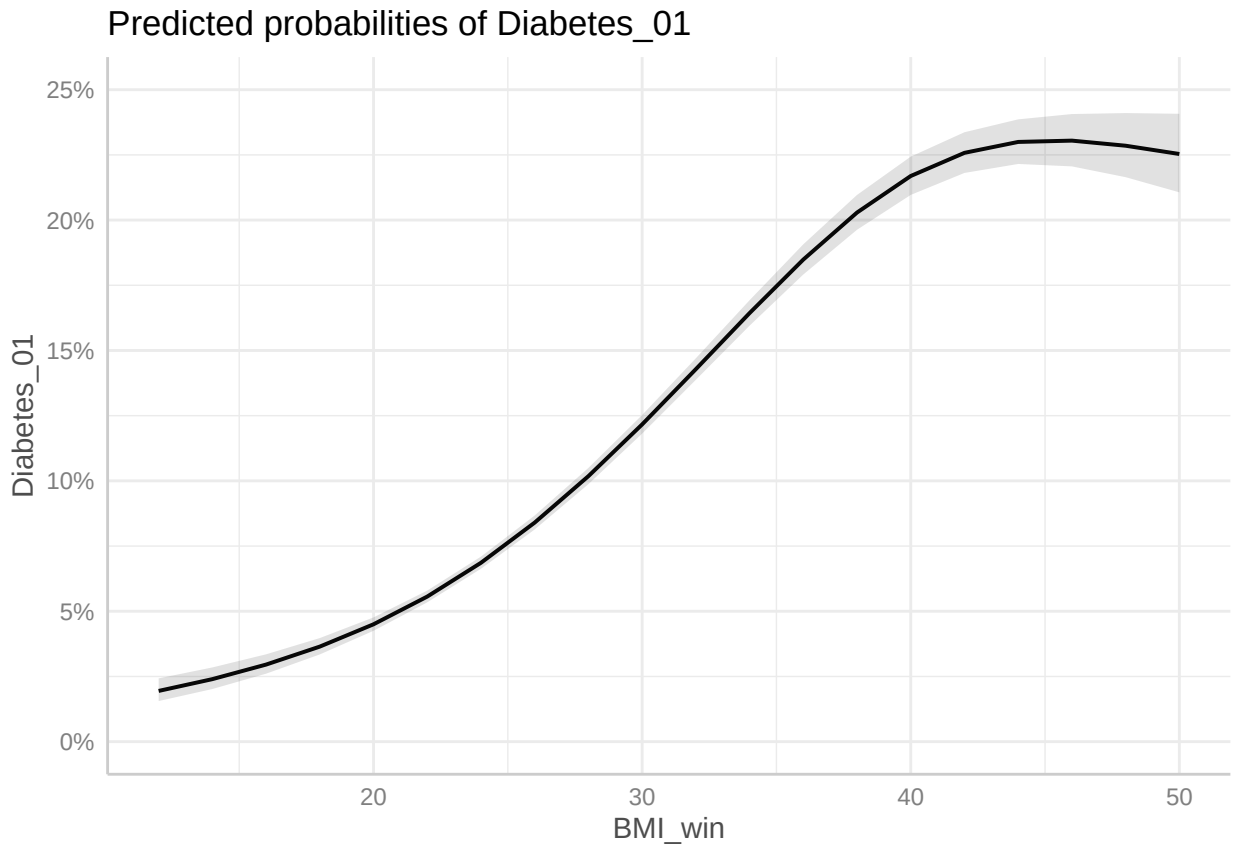
```
# AIC of the model without the non-significant variables
AIC(NoSignificant)
```

```
## [1] 132240.1
```

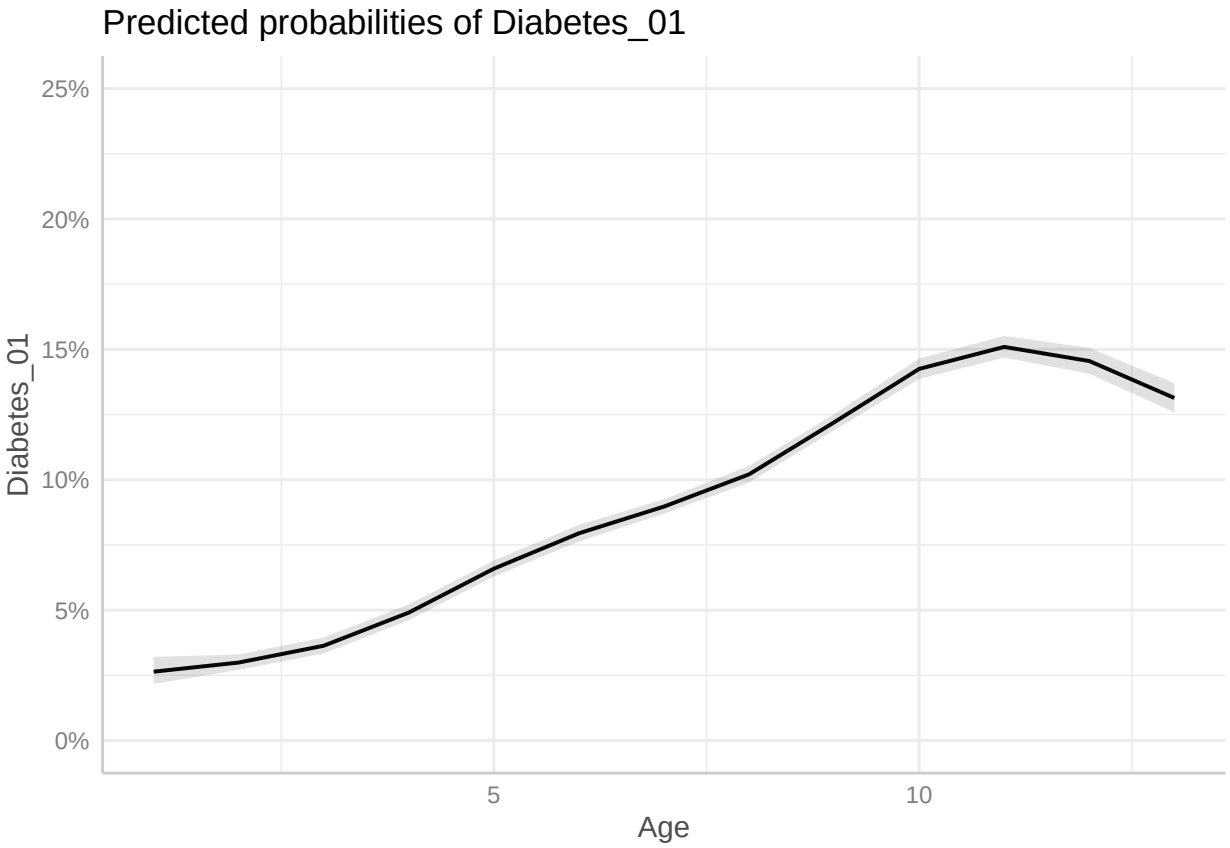
Looking at the above combinations, the full model has the lowest AIC so all variables were kept in the model. Individual variables were also tested but AIC values did not suggest removing any of the individual variables either.

Marginal Effects

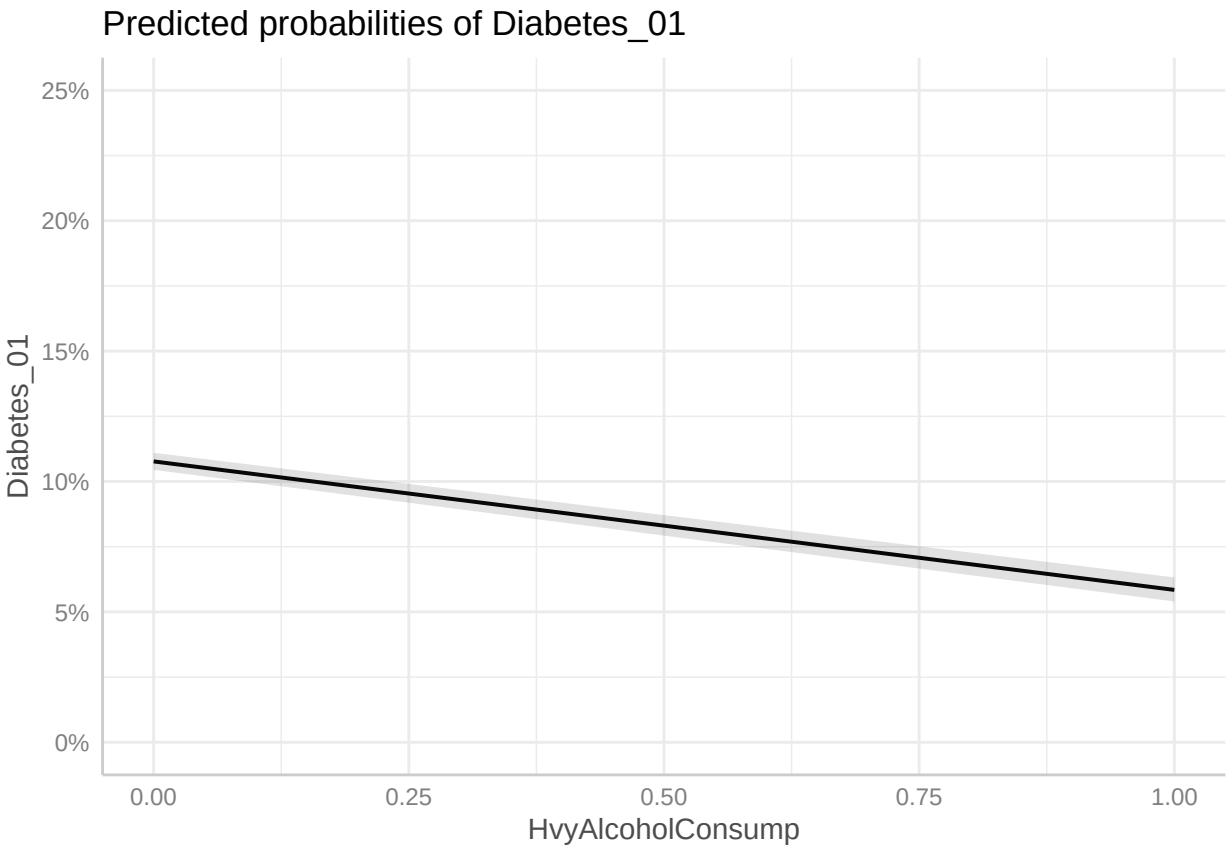
The below plots look at the effect of an attribute by varying that attribute while holding the rest of the variables constant. These are particularly informative for helping understand the terms where splines were applied (BMI, Age). I have also included Alcohol Consumption in this section because the relationship is not what I expected and deserves some explanation. The rest are included in the appendix because the relationships were straightforward.



Holding the other model variables constant, the effect of increasing BMI accelerates around 20 with the steepest part of the curve around 30 and flattens out around 45.



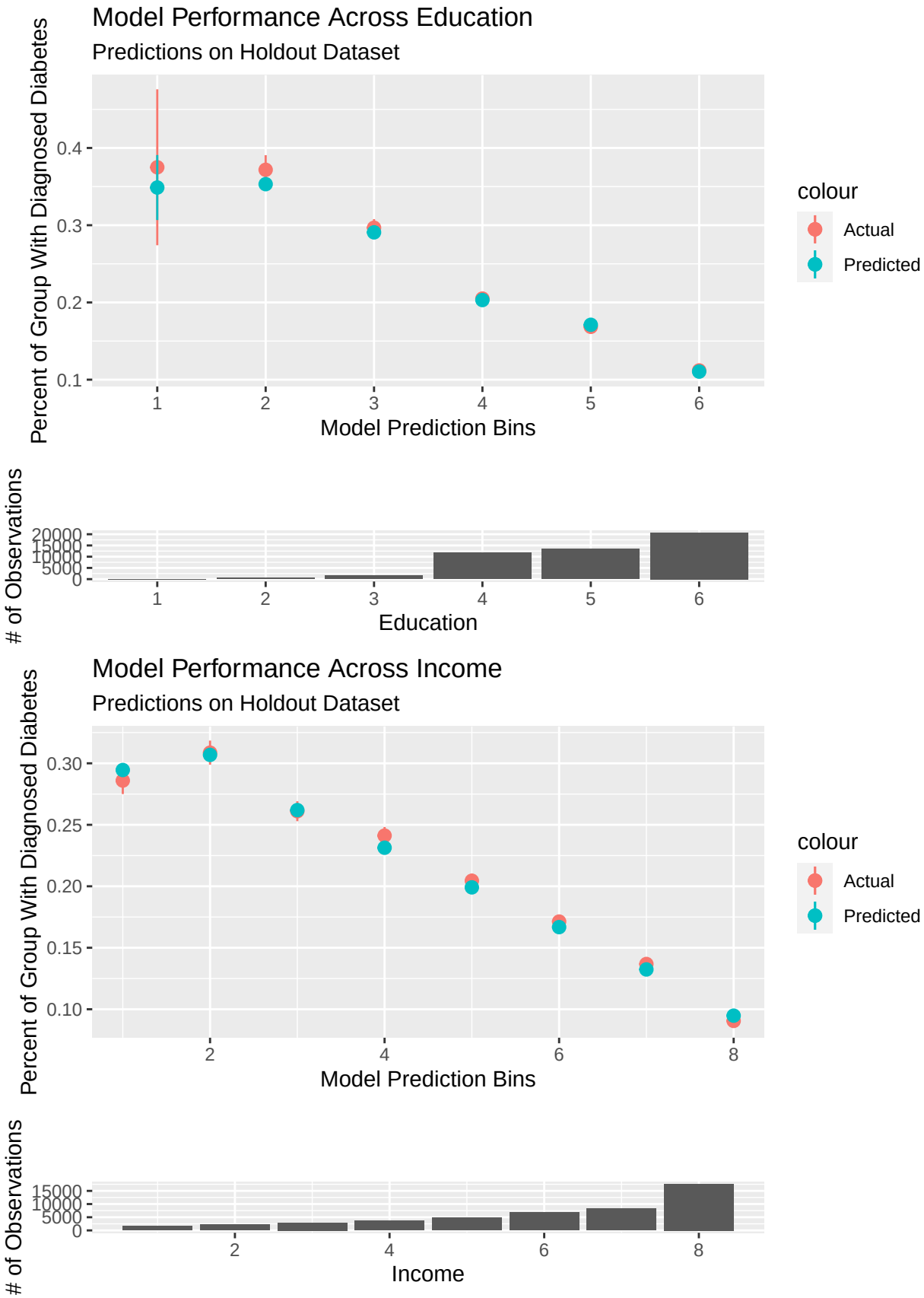
For Age, the effect is somewhat similar. It is relatively flat from 18-34 (1-3) with an increase in slope from 35 - 69 (4-10), but it has a dip towards the higher age ranges (11-13).

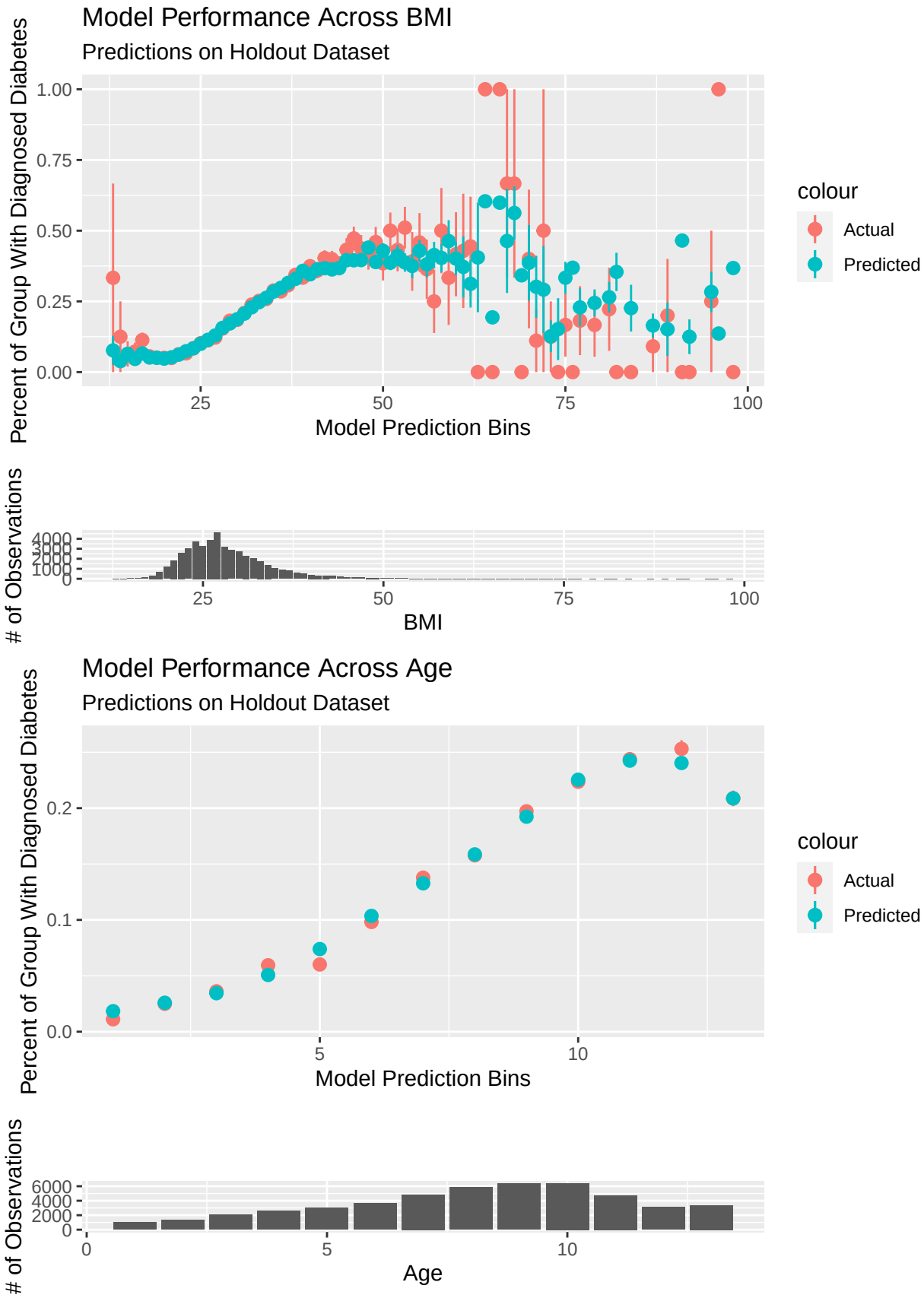


An interesting finding coming out of this was that Heavy Alcohol Consumption had a negative coefficient, meaning if a respondent was a heavy consumer of alcohol they were far less likely to have been diagnosed with diabetes. There is some research to suggest that alcohol may have a role in staving off diabetes (7), (8), (9), (10). Additionally, heavy alcohol consumption tends to exacerbate the symptoms of diabetes once an individual has it (11), (12), (13). To be clear, none of this analysis is medical advice and does not advise that anyone heavily consuming to lower the risk of diabetes.

Performance by Variable

Finally, we will now look at the residuals by variable. These charts help understand how well the signal of a particular variable is being captured. If the average predicted value is close to the average actual value, the model is utilizing that variable well. If there is a gap and there is a substantial number of records at that level, that suggest further transformation could be done to improve the model fit. Education, Income, BMI, and Age have been included here. The rest are included in the appendix.


Model Performance Across Education
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.4
0.3
0.2
0.1
1
2
3
4
5
6
Model Prediction Bins
colour
Actual
Predicted
# of Observations
20000
15000
10000
5000
0
Education
Model Performance Across Income
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.30
0.25
0.20
0.15
0.10
2
4
6
8
Model Prediction Bins
colour
Actual
Predicted
# of Observations
15000
10000
5000
0
Income


Model Performance Across BMI
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
1.00
0.75
0.50
0.25
0.00
25
50
75
100
Model Prediction Bins
colour
Actual
Predicted
# of Observations
4000
3000
2000
1000
0
25
50
75
100
BMI
Model Performance Across Age
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.2
0.1
0.0
5
10
Model Prediction Bins
colour
Actual
Predicted
# of Observations
6000
4000
2000
0
0
5
10
Age

As seen above, anywhere with substantial numbers of observations the model fit is reasonably good. For extreme values of BMI, there is some noise because there are so few observations in each level.

Conclusion

In summary, a model was built to predict diabetes with the intent of using the model to help identify cases of undiagnosed diabetes as well as individuals who could benefit from intervention and coaching. While there is no direct way to measure the performance of the model in predicting *undiagnosed* cases of diabetes, the model does perform reasonably well at predicting cases of diagnosed diabetes. It is reasonable to believe that cases of undiagnosed diabetes would also score highly in the model. HealthInsure Co. can use this model on their data to identify individuals at high risk for developing diabetes or having undiagnosed diabetes to effectively empower members to make informed decisions about their health.

Future Work

The next steps would be to pilot an outreach program using this model and analyze the real world results. From there, further analysis could be done to look at respondents over time. This may help identify individuals on the path towards diabetes instead of ones that are already at high risk. Additionally, there may be other data elements with predictive power or further interactions not captured by this current model.

References

(1) Centers for Disease Control and Prevention. (2021, December 29). Prevalence of both diagnosed and undiagnosed diabetes. Centers for Disease Control and Prevention. Retrieved August 1, 2022, from https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html

(2) The cost of diabetes. The Cost of Diabetes | ADA. (n.d.). Retrieved August 1, 2022, from https://www.diabetes.org/about-us/statistics/cost-diabetes

(3) Teboul, A. (2021, November 8). Diabetes health indicators dataset. Kaggle. Retrieved August 1, 2022, from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

(4) Alexteboul. (2022, March 10). Diabetes health indicators dataset notebook. Kaggle. Retrieved July 1, 2022, from https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook/notebook

(5) Hastie, T. J. (1992) Generalized additive models. Chapter 7 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

(6) Centers for Disease Control and Prevention. (2017, August 11). CDC - 2015 BRFSS survey data and Documentation. Centers for Disease Control and Prevention. Retrieved August 1, 2022, from https://www.cdc.gov/brfss/annual_data/annual_2015.html

(7) Craig Knott, Steven Bell, Annie Britton; Alcohol Consumption and the Risk of Type 2 Diabetes: A Systematic Review and Dose-Response Meta-analysis of More Than 1.9 Million Individuals From 38 Observational Studies. Diabetes Care 1 September 2015; 38 (9): 1804–1812. https://doi.org/10.2337/dc15-0710

(8) Koppes LL, Dekker JM, Hendriks HF, Bouter LM, Heine RJ. Moderate alcohol consumption lowers the risk of type 2 diabetes: a meta-analysis of prospective observational studies. Diabetes Care. 2005 Mar;28(3):719-25. doi: 10.2337/diacare.28.3.719. PMID: 15735217.

(9) Shai I, Wainstein J, Harman-Boehm I, Raz I, Fraser D, Rudich A, Stampfer MJ. Glycemic effects of moderate alcohol intake among patients with type 2 diabetes: a multicenter, randomized, clinical intervention trial. Diabetes Care. 2007 Dec;30(12):3011-6. doi: 10.2337/dc07-1103. Epub 2007 Sep 11. PMID: 17848609.

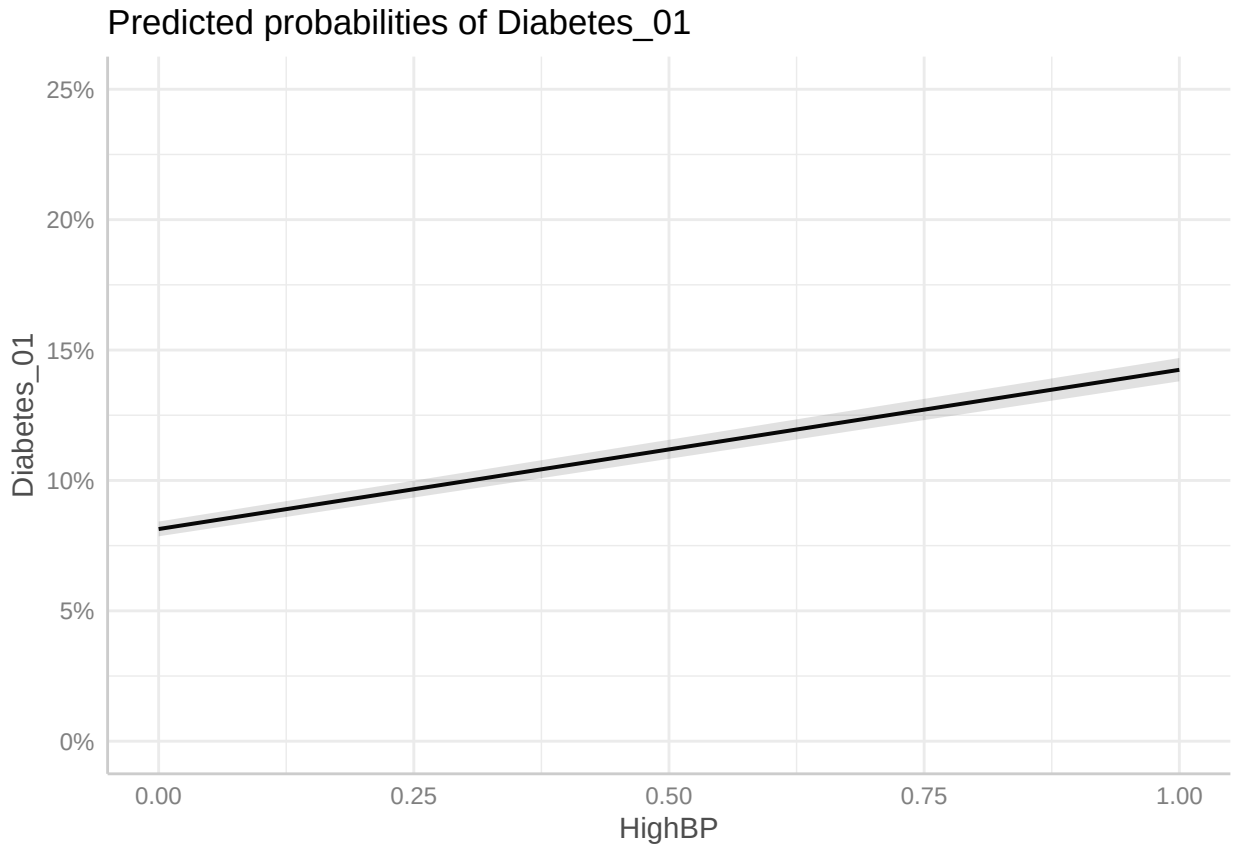
(10) Brand-Miller JC, Fatema K, Middlemiss C, Bare M, Liu V, Atkinson F, Petocz P. Effect of alcoholic beverages on postprandial glycemia and insulinemia in lean, young, healthy adults. Am J Clin Nutr. 2007 Jun;85(6):1545-51. doi: 10.1093/ajcn/85.6.1545. Erratum in: Am J Clin Nutr. 2007 Sep;86(3):808. Fatima, Kaniz [corrected to Fatema, Kaniz]. PMID: 17556691.

(11) McCulloch DK, Campbell IW, Prescott RJ, Clarke BF. Effect of alcohol intake on symptomatic peripheral neuropathy in diabetic men. Diabetes Care. 1980 Mar-Apr;3(2):245-7. doi: 10.2337/diacare.3.2.245. PMID: 7389544.
(12) O'Keefe SJ, Marks V. Lunchtime gin and tonic a cause of reactive hypoglycaemia. Lancet. 1977 Jun 18;1(8025):1286-8. doi: 10.1016/s0140-6736(77)91321-6. PMID: 68385.
(13) Richardson T, Weiss M, Thomas P, Kerr D. Day after the night before: influence of evening alcohol on risk of hypoglycemia in patients with type 1 diabetes. Diabetes Care. 2005 Jul;28(7):1801-2. doi: 10.2337/diacare.28.7.1801. PMID: 15983341.
(14) Complications of diabetes. Diabetes UK. Retrieved August 1, 2022, from https://www.diabetes.org.uk/guide-to-diabetes/complications

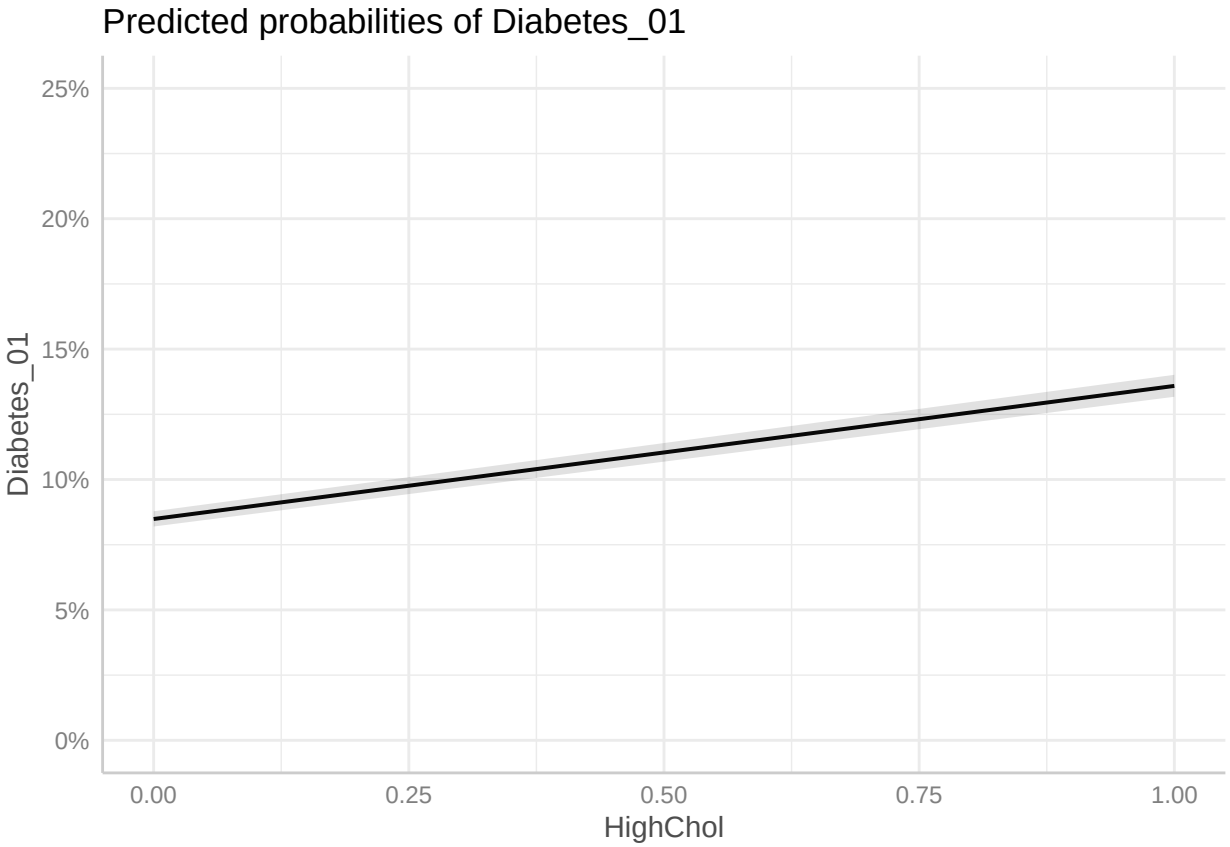
Appendix

Marginal Plots

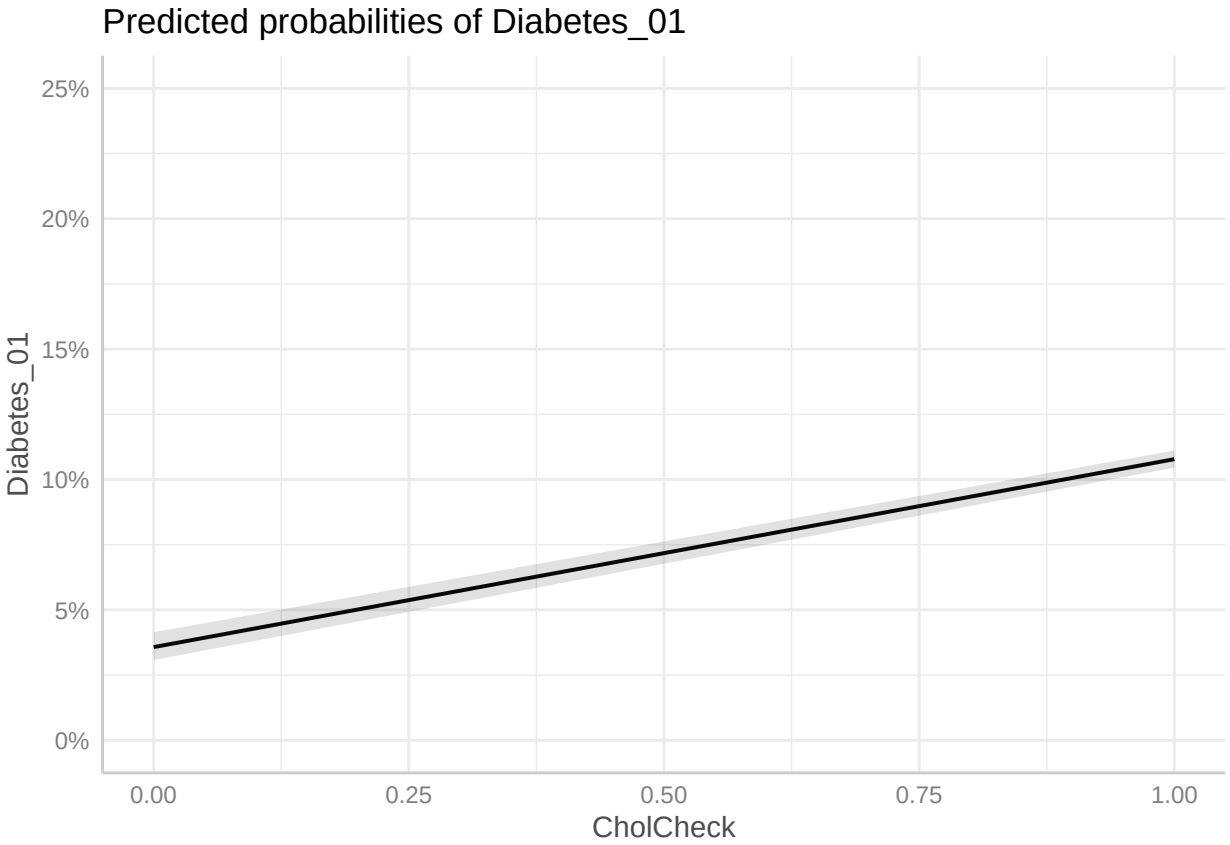
```
## [[1]]
```



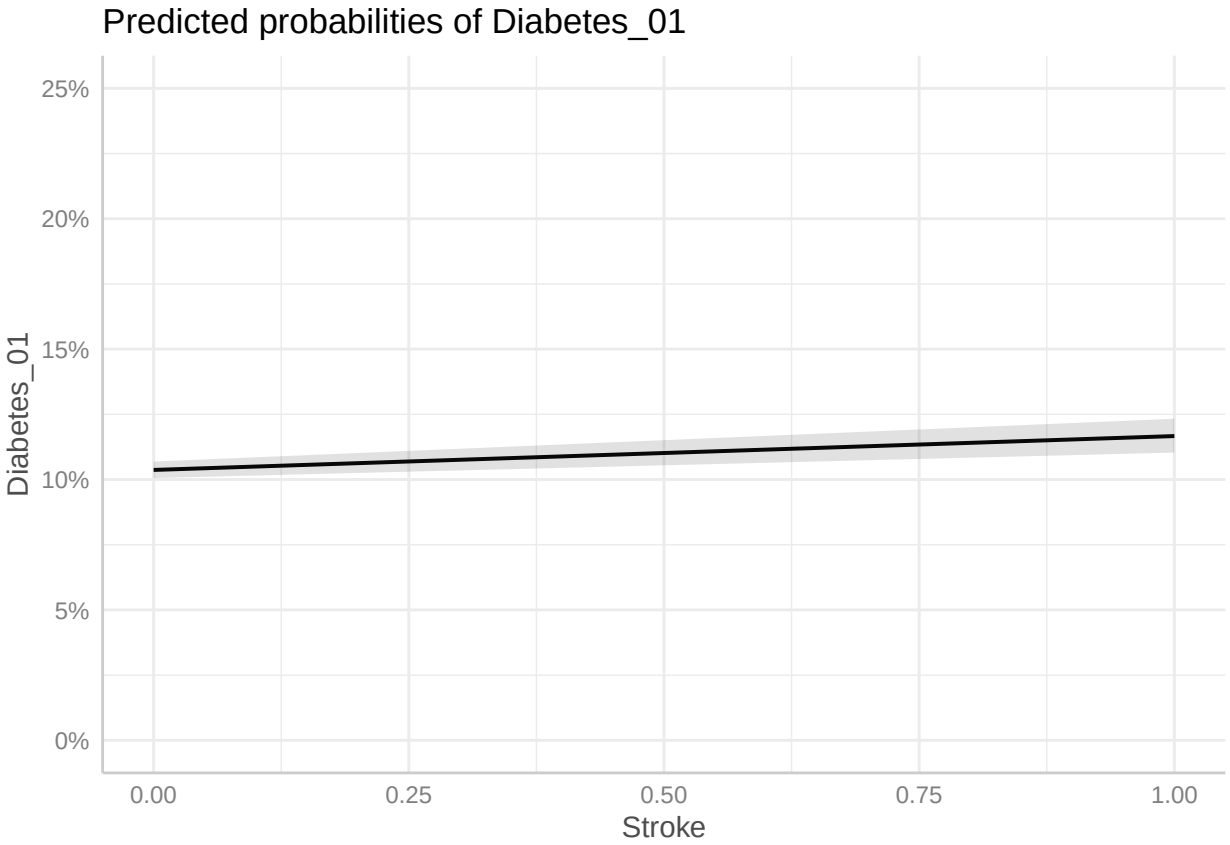

```
##
## [[2]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
0.00
0.25
0.50
0.75
1.00
HighChol

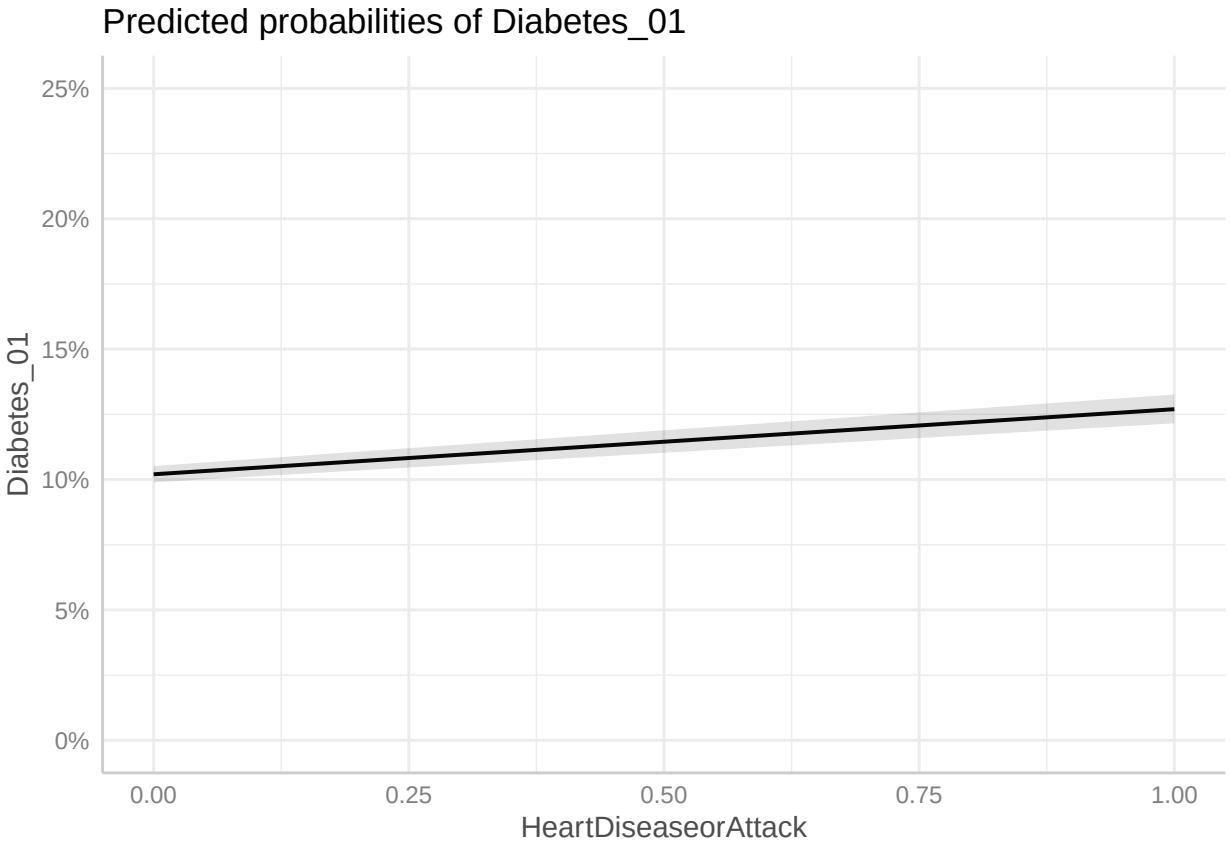
```
##
## [[3]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
0.00
0.25
0.50
0.75
1.00
CholCheck

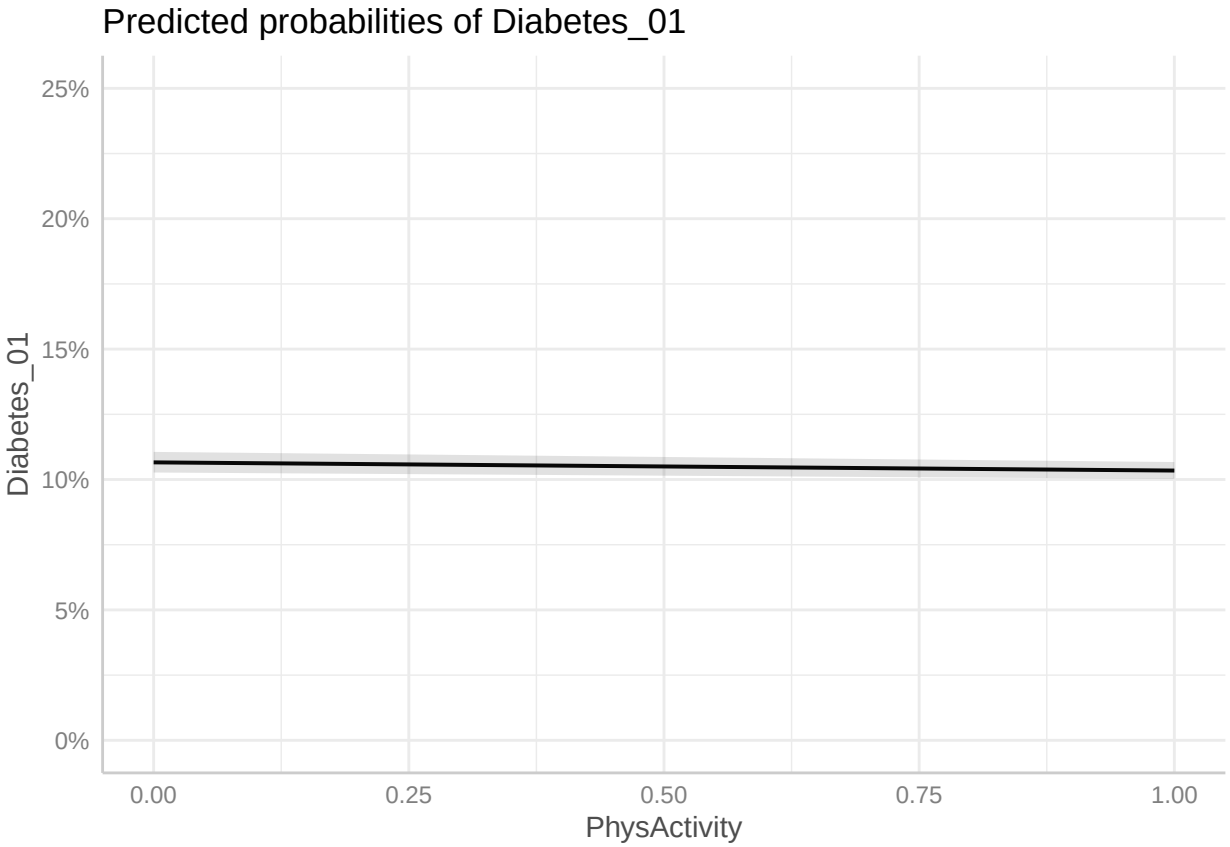
```
##
## [[4]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
0.00
0.25
0.50
0.75
1.00
Stroke

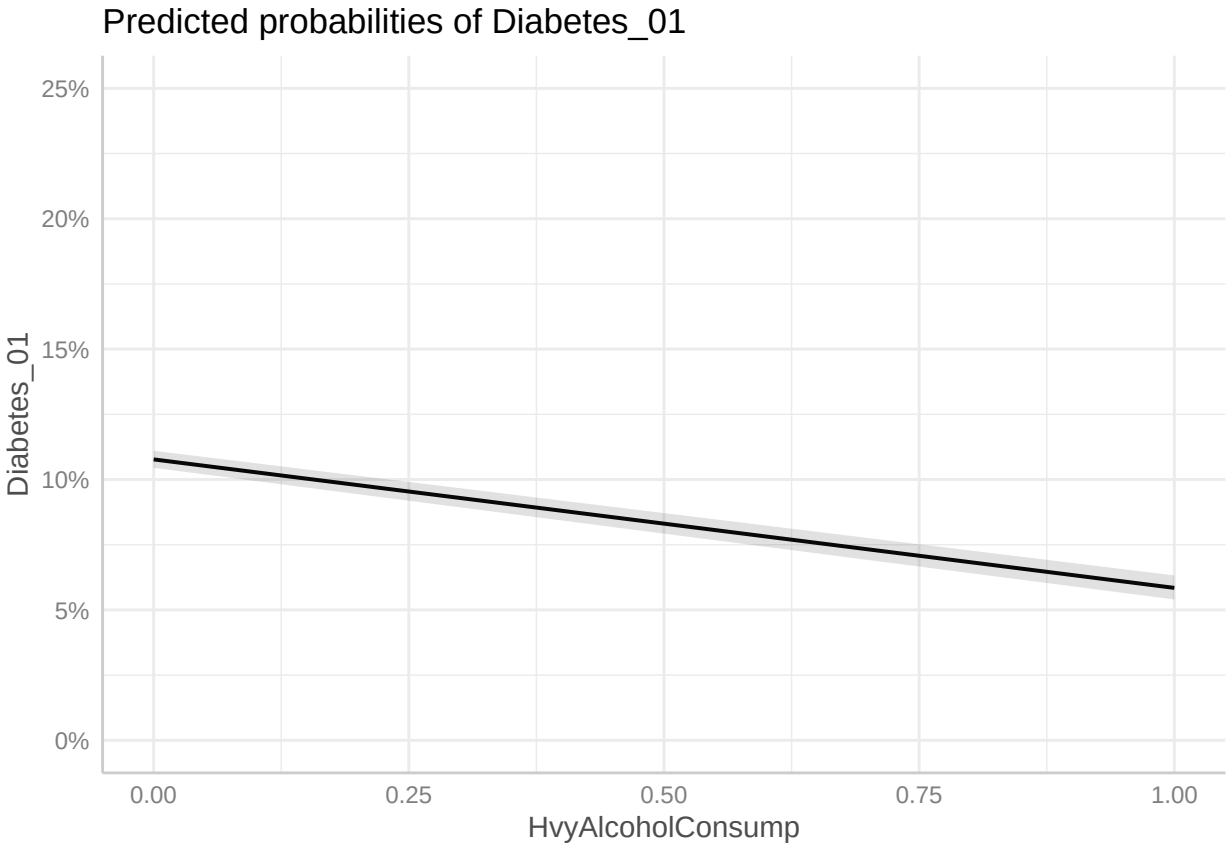
```
##
## [[5]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
0.00
0.25
0.50
0.75
1.00
HeartDiseaseorAttack

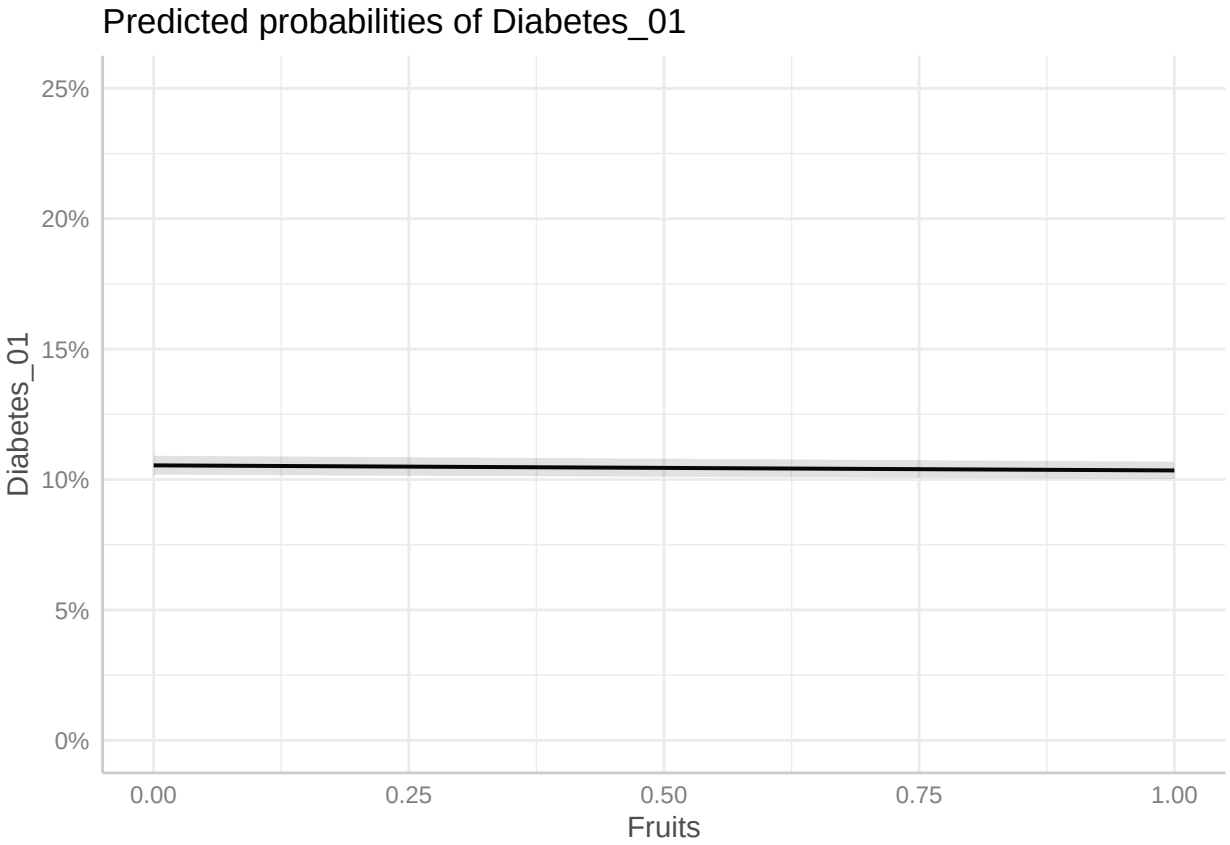
```
##
## [[6]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
0.00
0.25
0.50
0.75
1.00
PhysActivity

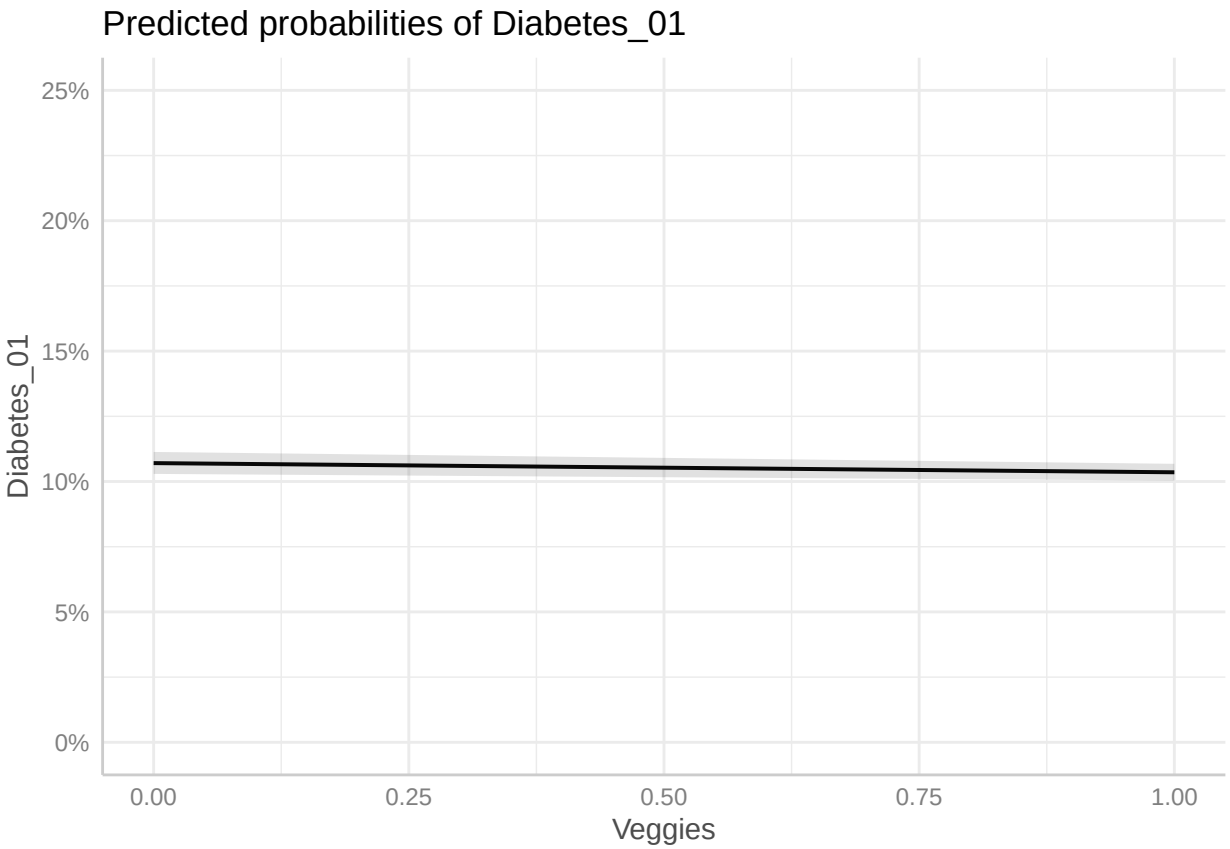
```
##
## [[7]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
0.00
0.25
0.50
0.75
1.00
HvyAlcoholConsump

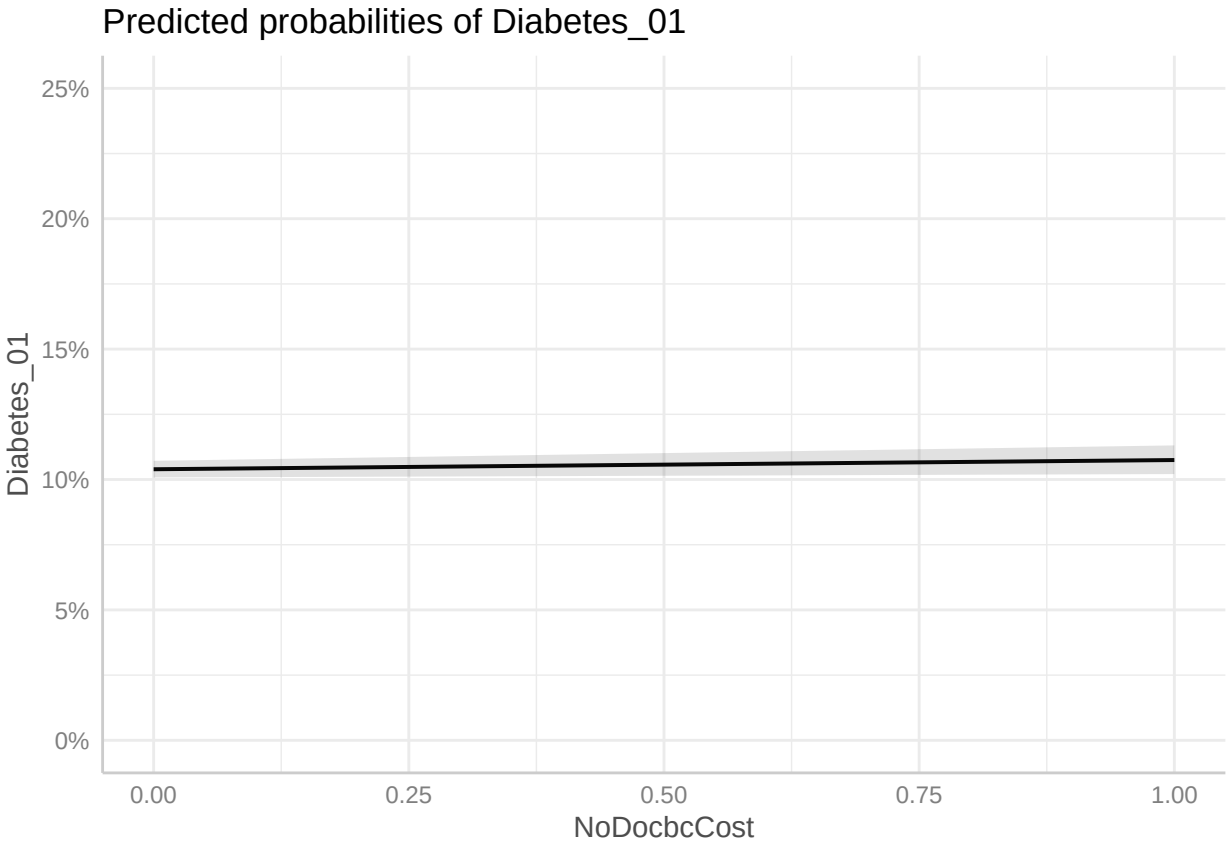
```
##
## [[8]]
```


Predicted probabilities of Diabetes_01
Diabetes_01
25%
20%
15%
10%
5%
0%
0.00
0.25
0.50
0.75
1.00
Fruits

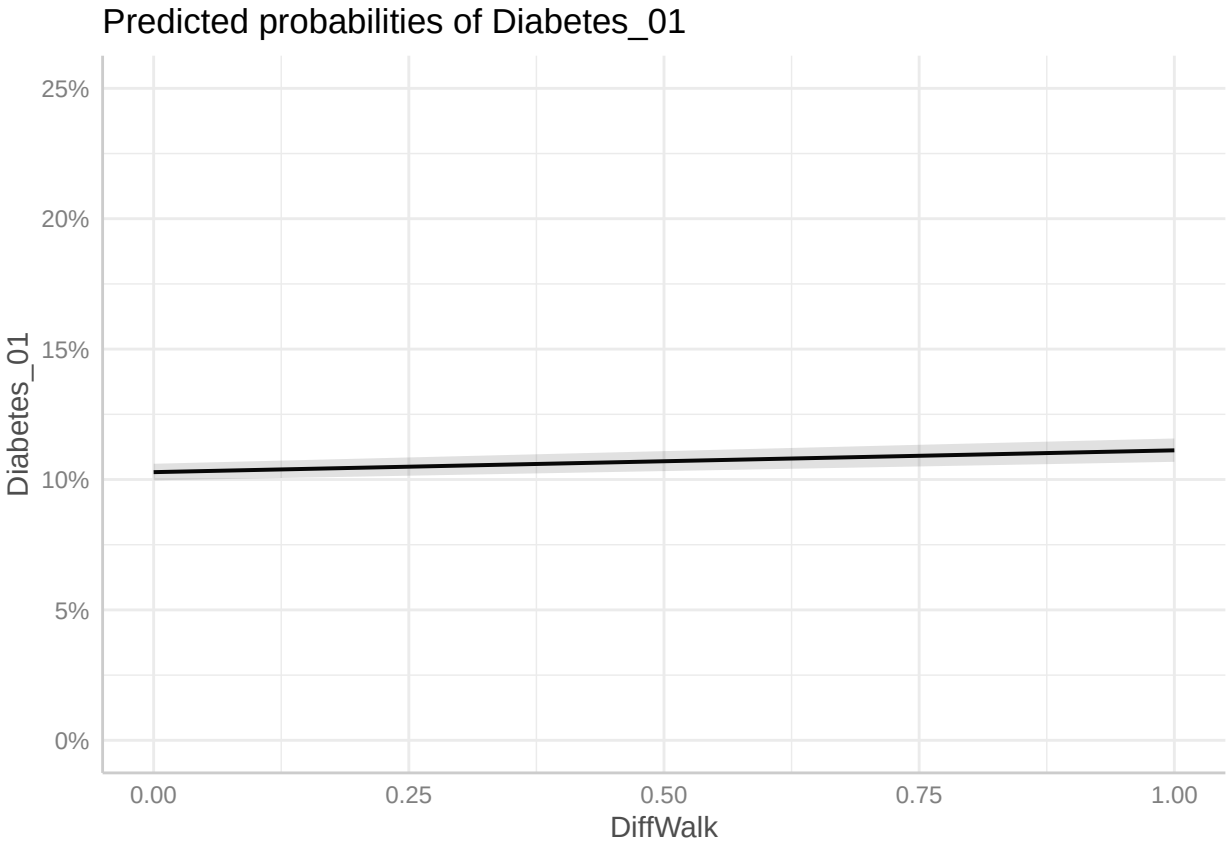
```
##
## [[9]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
0.00
0.25
0.50
0.75
1.00
Veggies

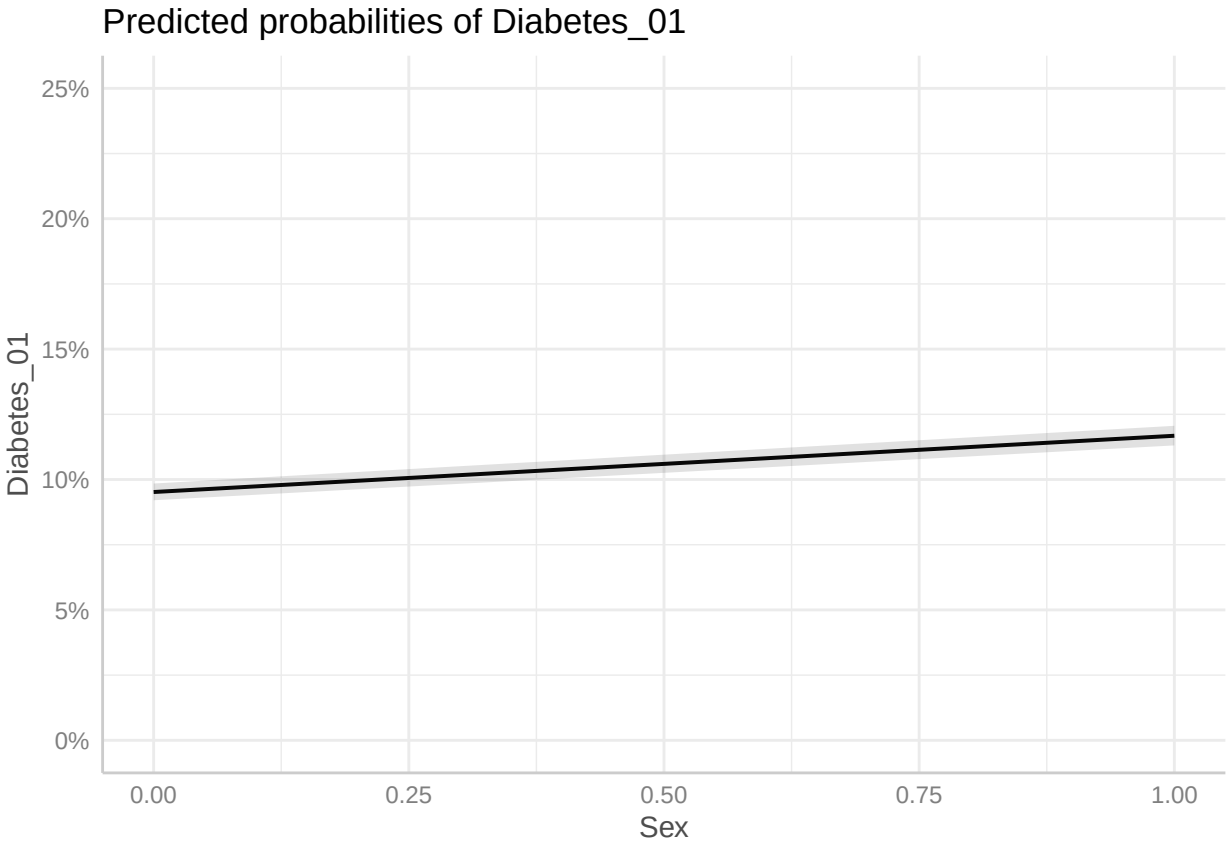
```
##
## [[10]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
0.00
0.25
0.50
0.75
1.00
NoDocbcCost

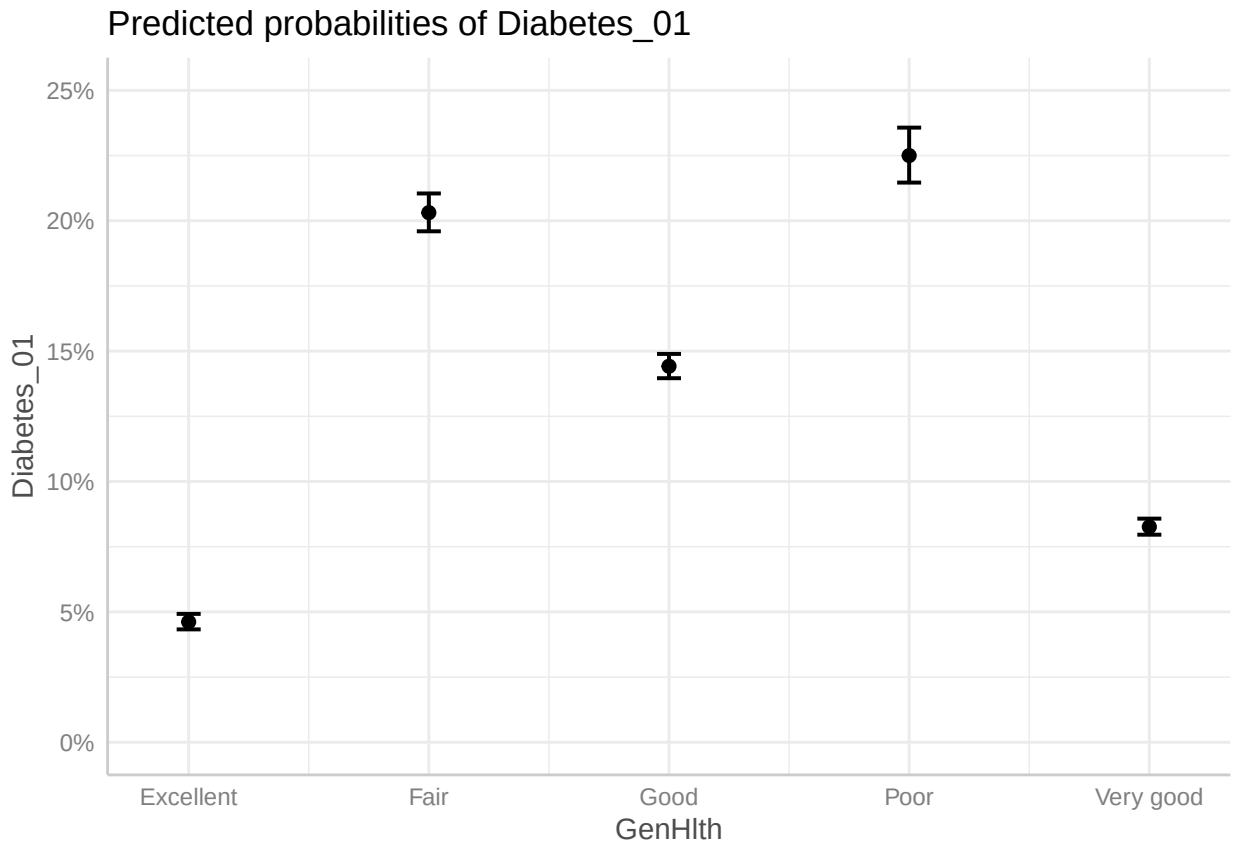
```
##
## [[11]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
0.00
0.25
0.50
0.75
1.00
DiffWalk

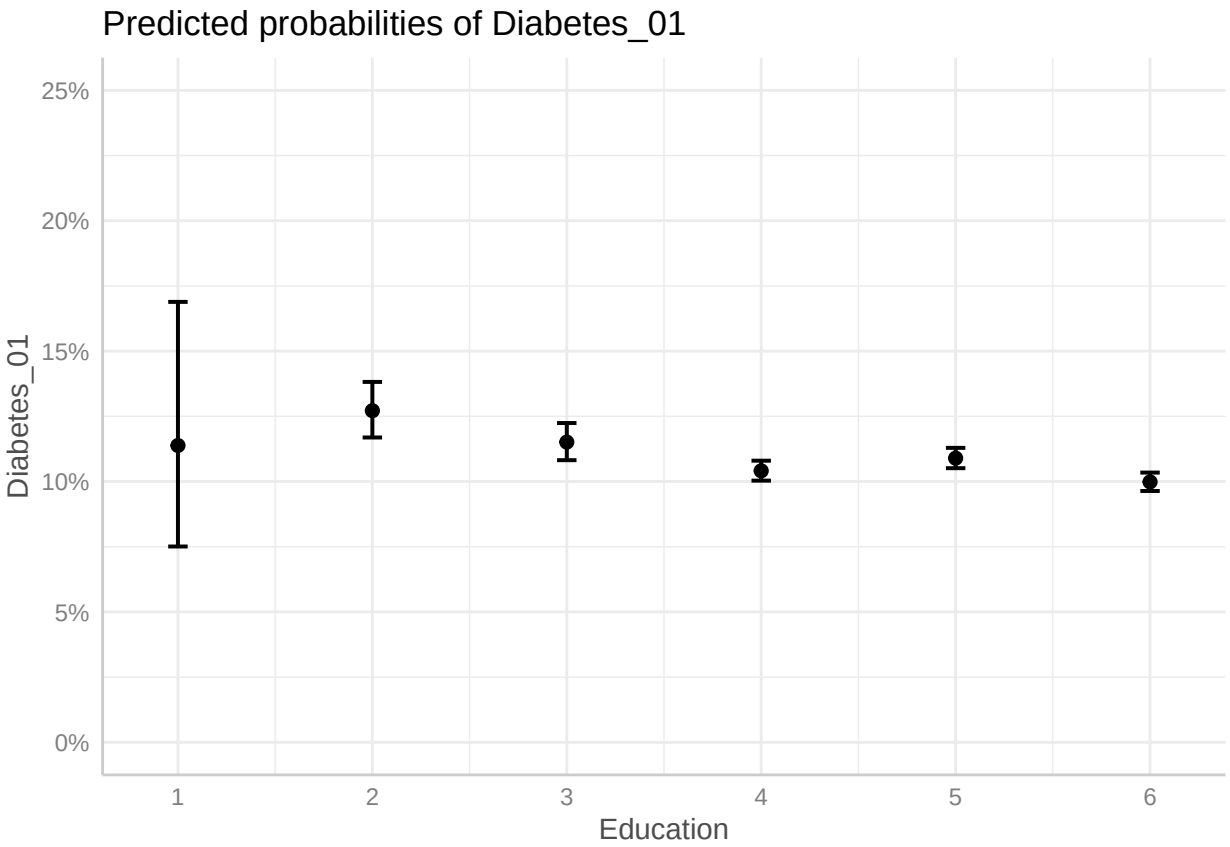
```
##
## [[12]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
0.00
0.25
0.50
0.75
1.00
Sex

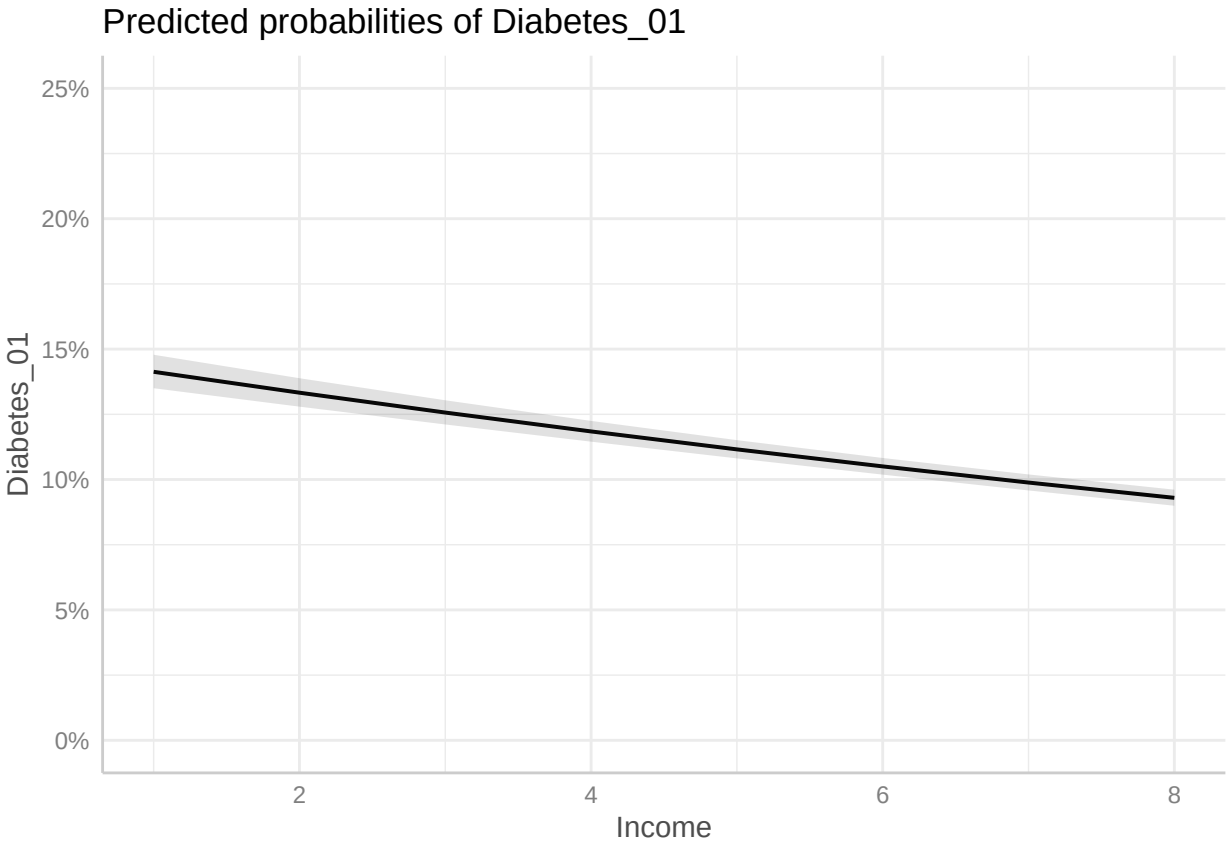
```
##
## [[13]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
Excellent
Fair
Good
Poor
Very good
GenHlth

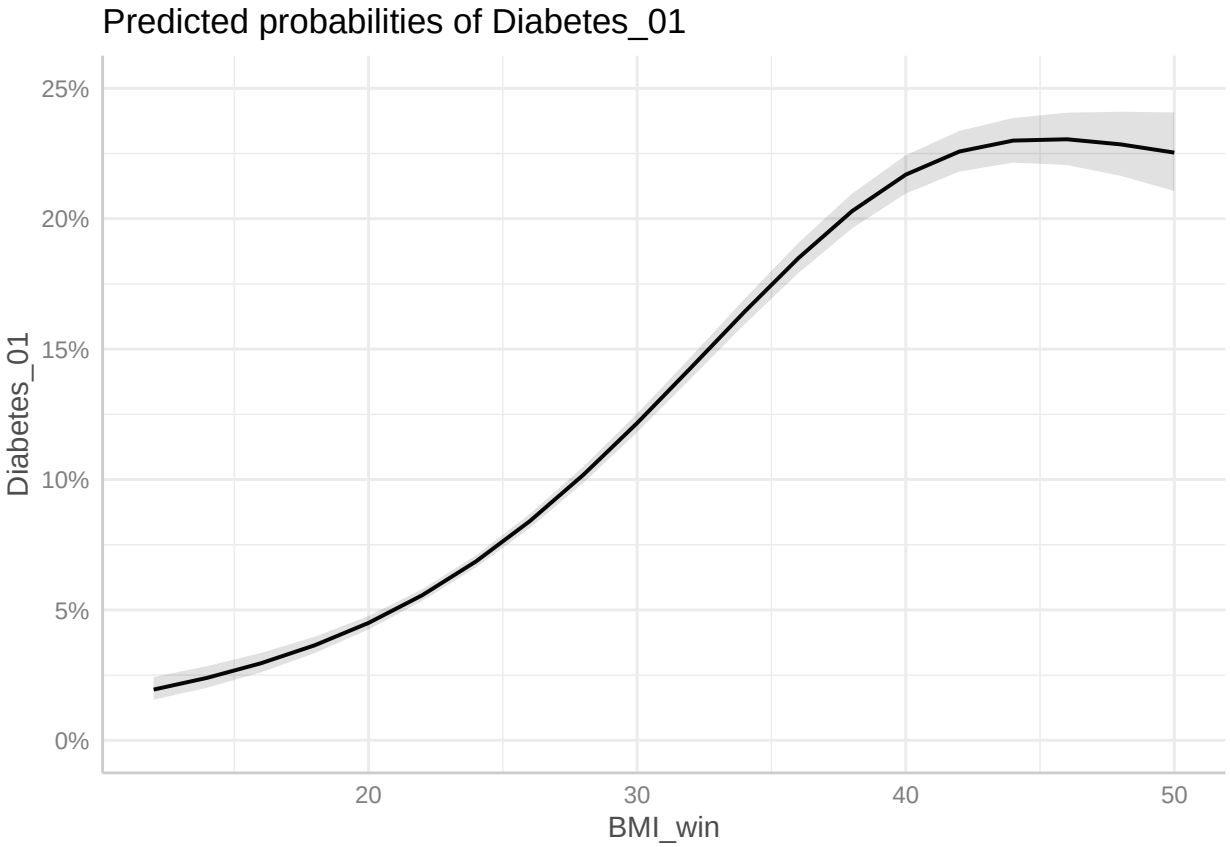
```
##
## [[14]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
1
2
3
4
5
6
Education

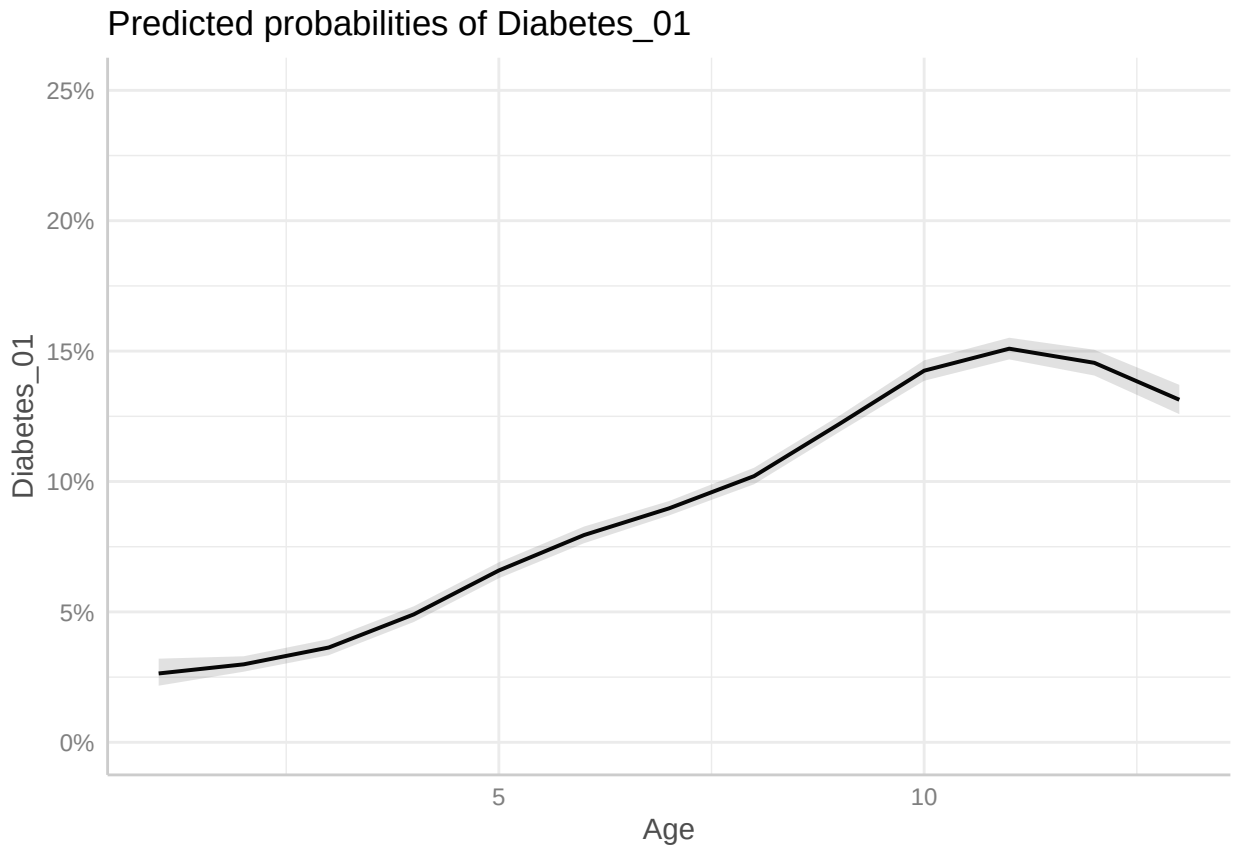
```
##
## [[15]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
2
4
6
8
Income

```
##
## [[16]]
```


Predicted probabilities of Diabetes_01
25%
20%
15%
10%
5%
0%
Diabetes_01
20
30
40
50
BMI_win

```
##
## [[17]]
```



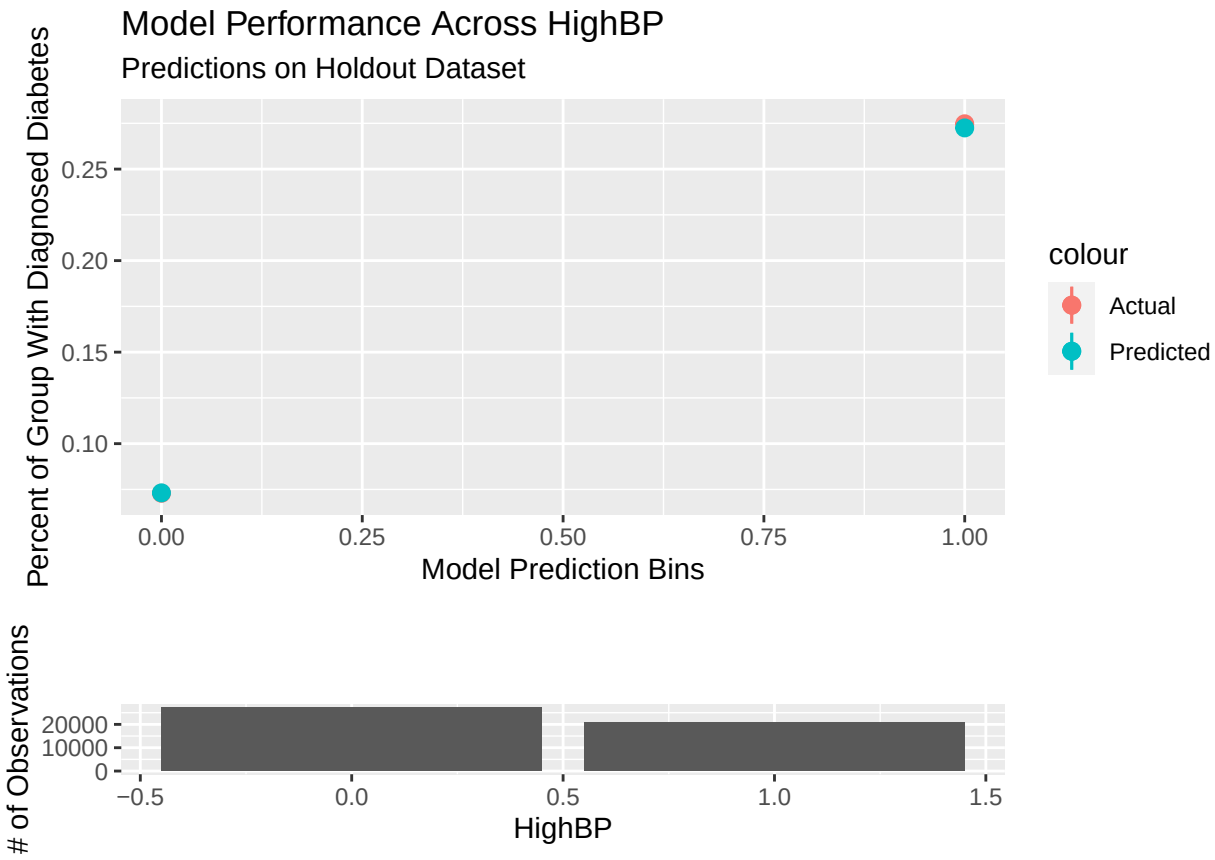
Confusion Matrix The confusion matrix is being included using Youden's J statistic for the cutoff value.

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    DiabetesNo DiabetesYes
##   DiabetesNo       28232        1495
##   DiabetesYes      12284        6241
##
##                Accuracy : 0.7144
##                  95% CI : (0.7104, 0.7185)
##     No Information Rate : 0.8397
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3219
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6968
##             Specificity : 0.8067
##          Pos Pred Value : 0.9497
##          Neg Pred Value : 0.3369
##              Prevalence : 0.8397
##          Detection Rate : 0.5851
##    Detection Prevalence : 0.6161
```

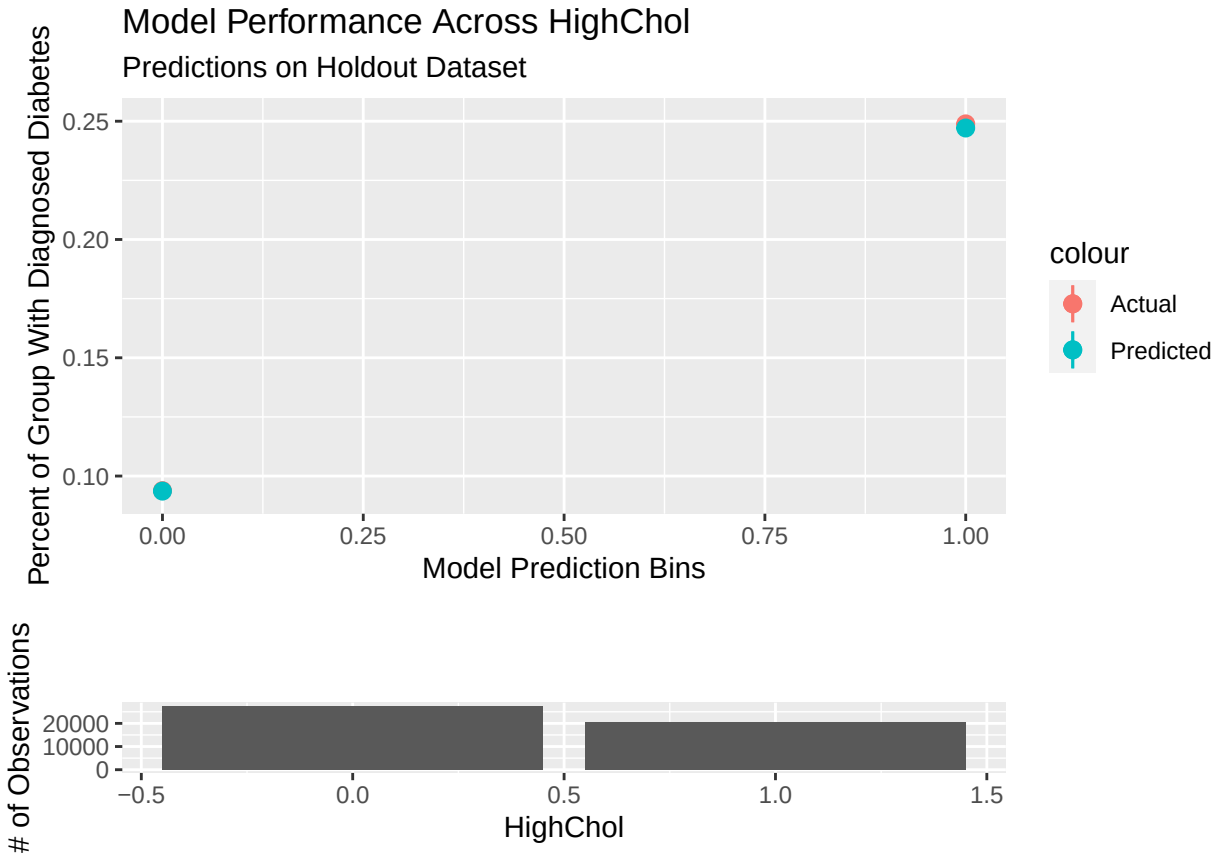
```
##        Balanced Accuracy : 0.7518
##
##         'Positive' Class : DiabetesNo
##
```

Residual Plots

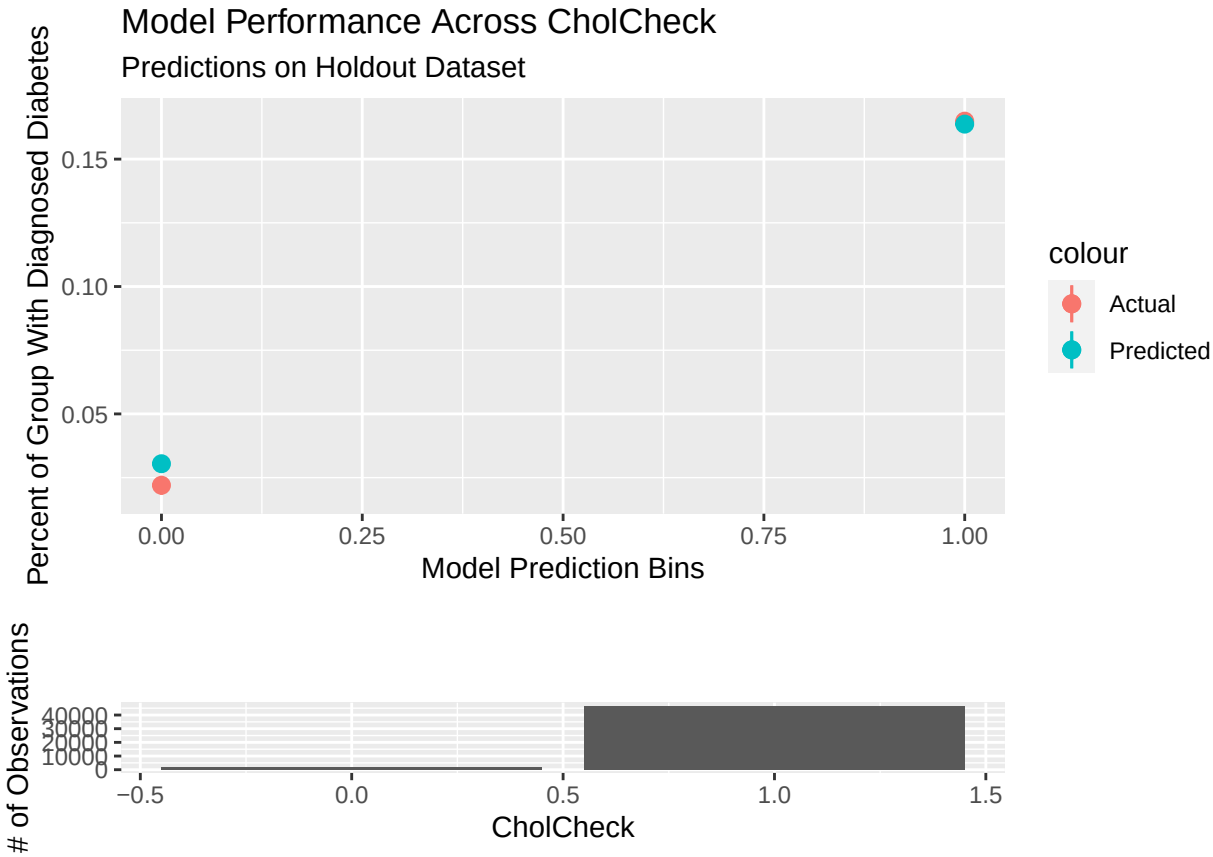
```
## [[1]]
```



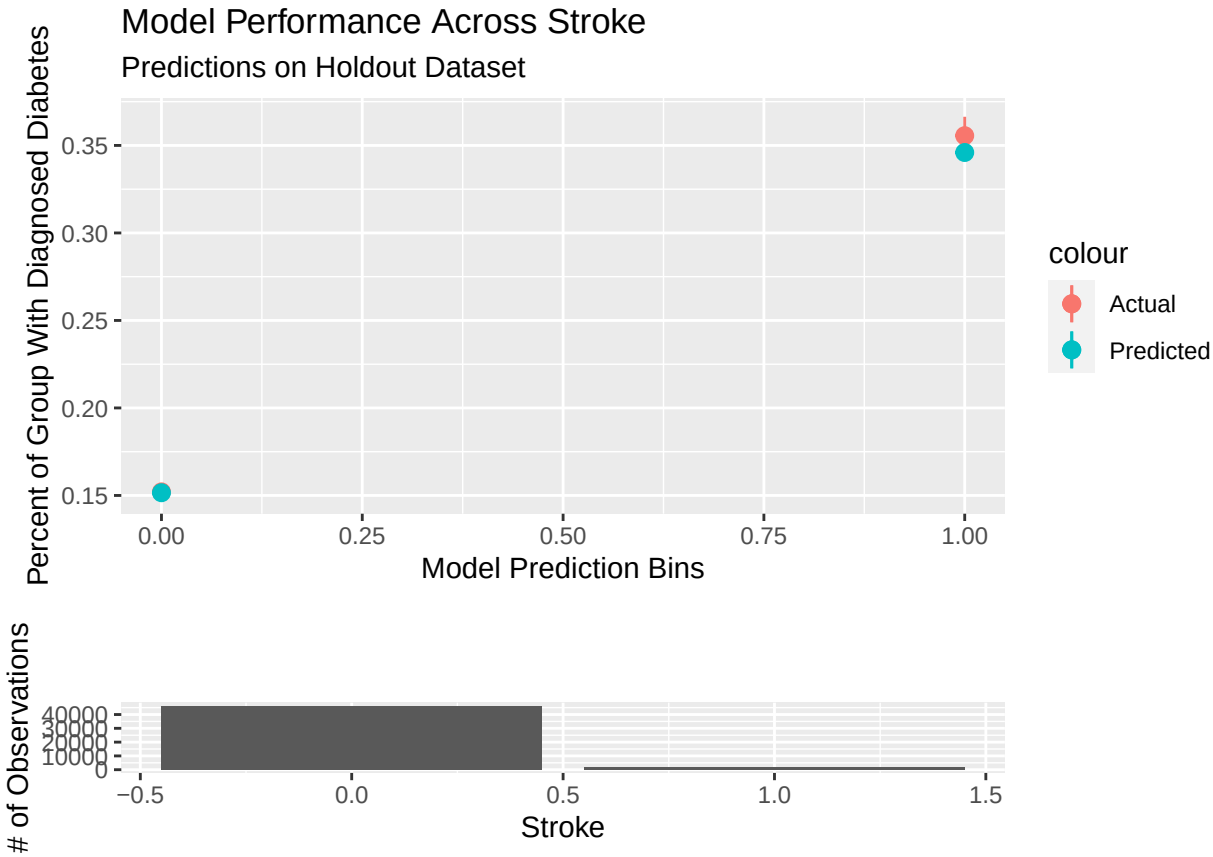

```
##
## [[2]]
```


Model Performance Across HighChol
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.25
0.20
0.15
0.10
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
HighChol

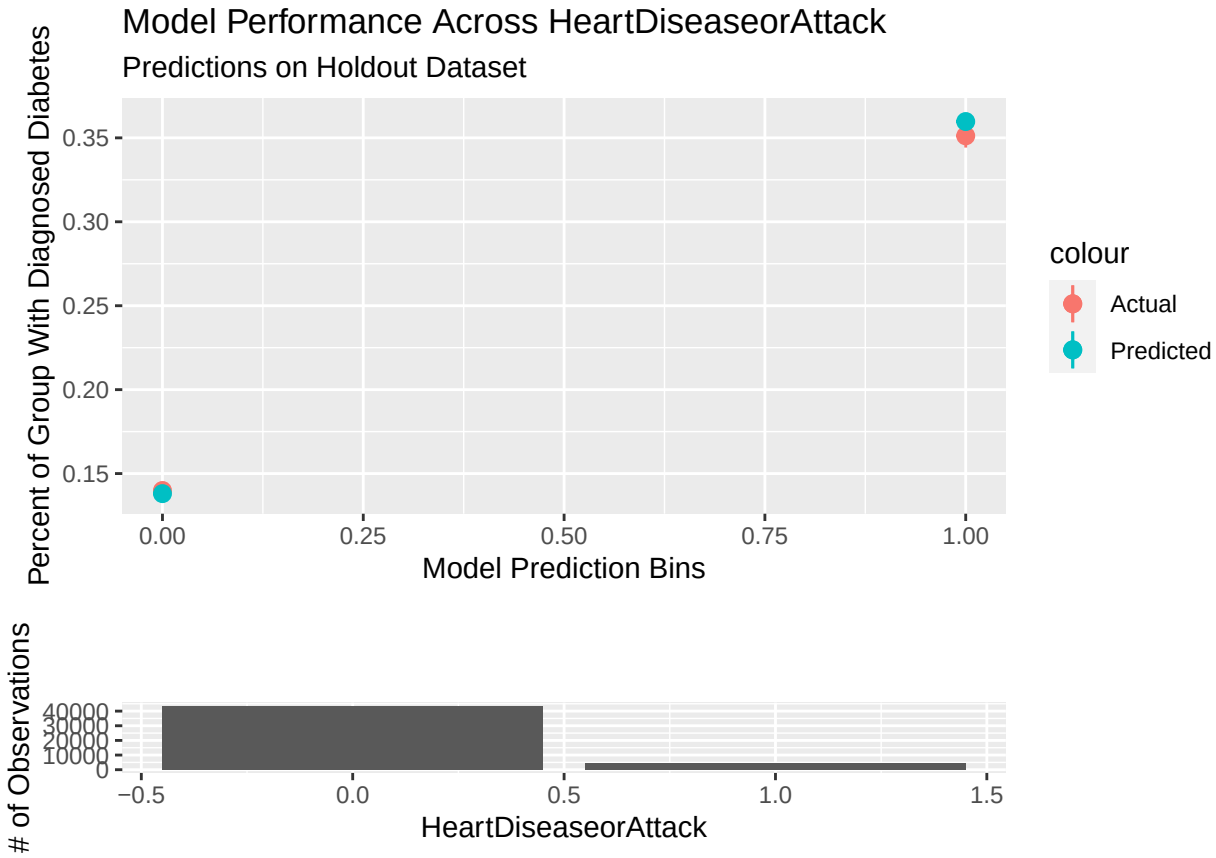
```
##
## [[3]]
```


Model Performance Across CholCheck
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.15
0.10
0.05
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
40000
30000
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
CholCheck

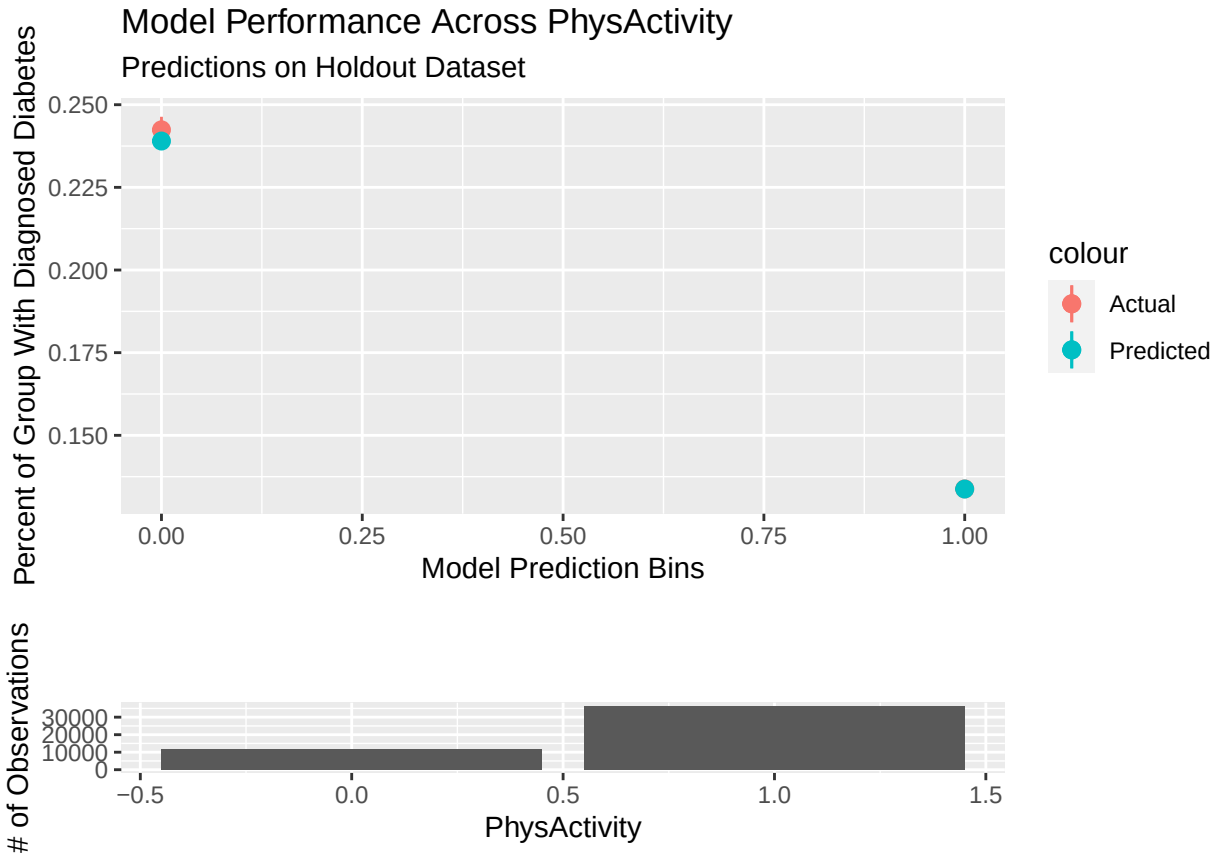
```
##
## [[4]]
```


Model Performance Across Stroke
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.35
0.30
0.25
0.20
0.15
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
40000
30000
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
Stroke

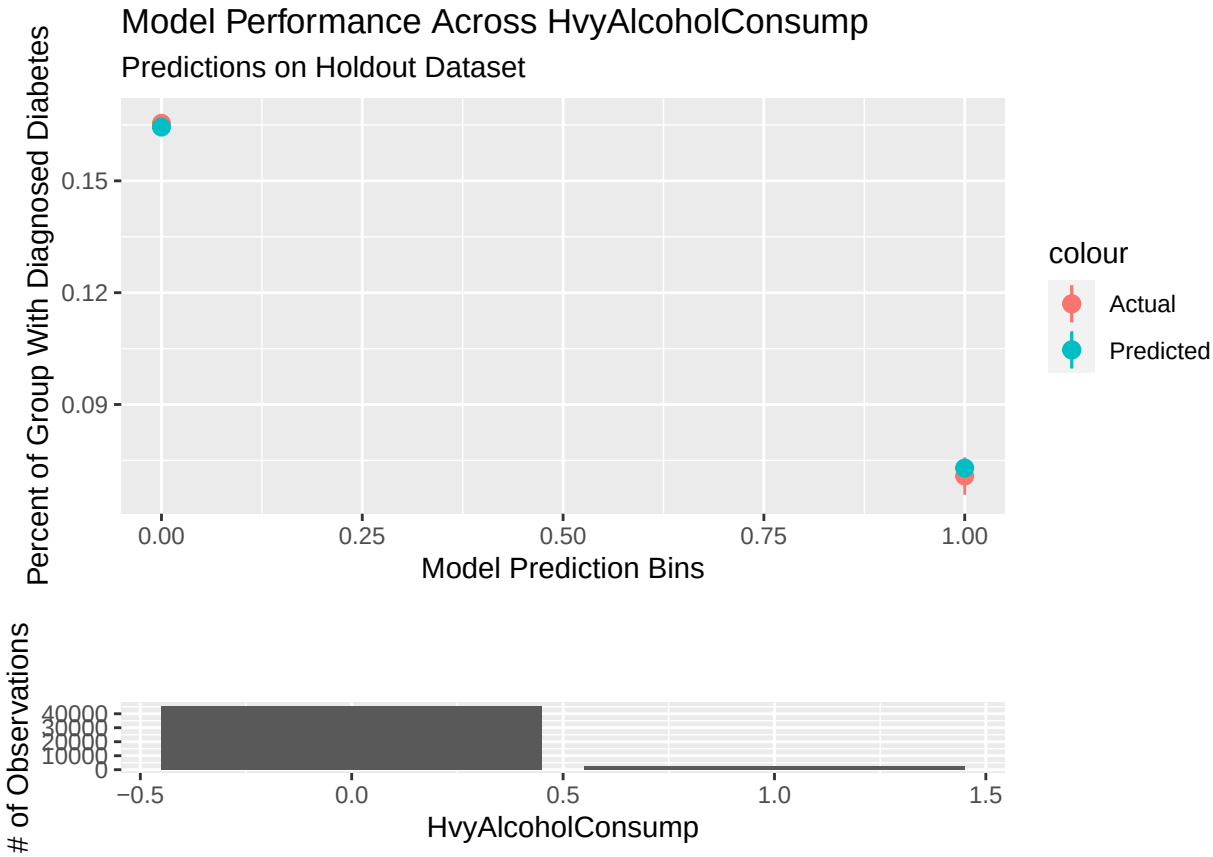
```
##
## [[5]]
```


Model Performance Across HeartDiseaseorAttack
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.35
0.30
0.25
0.20
0.15
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
40000
30000
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
HeartDiseaseorAttack

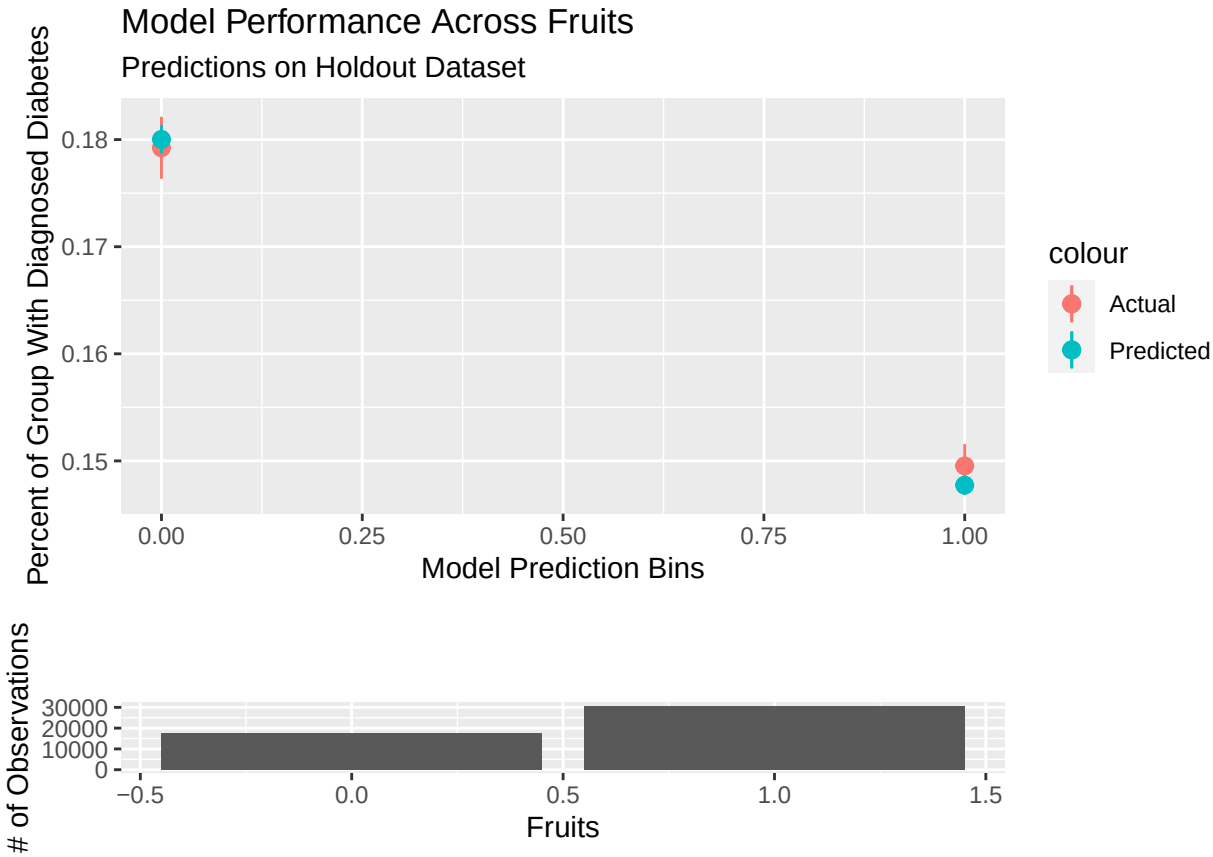
```
##
## [[6]]
```


Model Performance Across PhysActivity
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.250
0.225
0.200
0.175
0.150
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
30000
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
PhysActivity

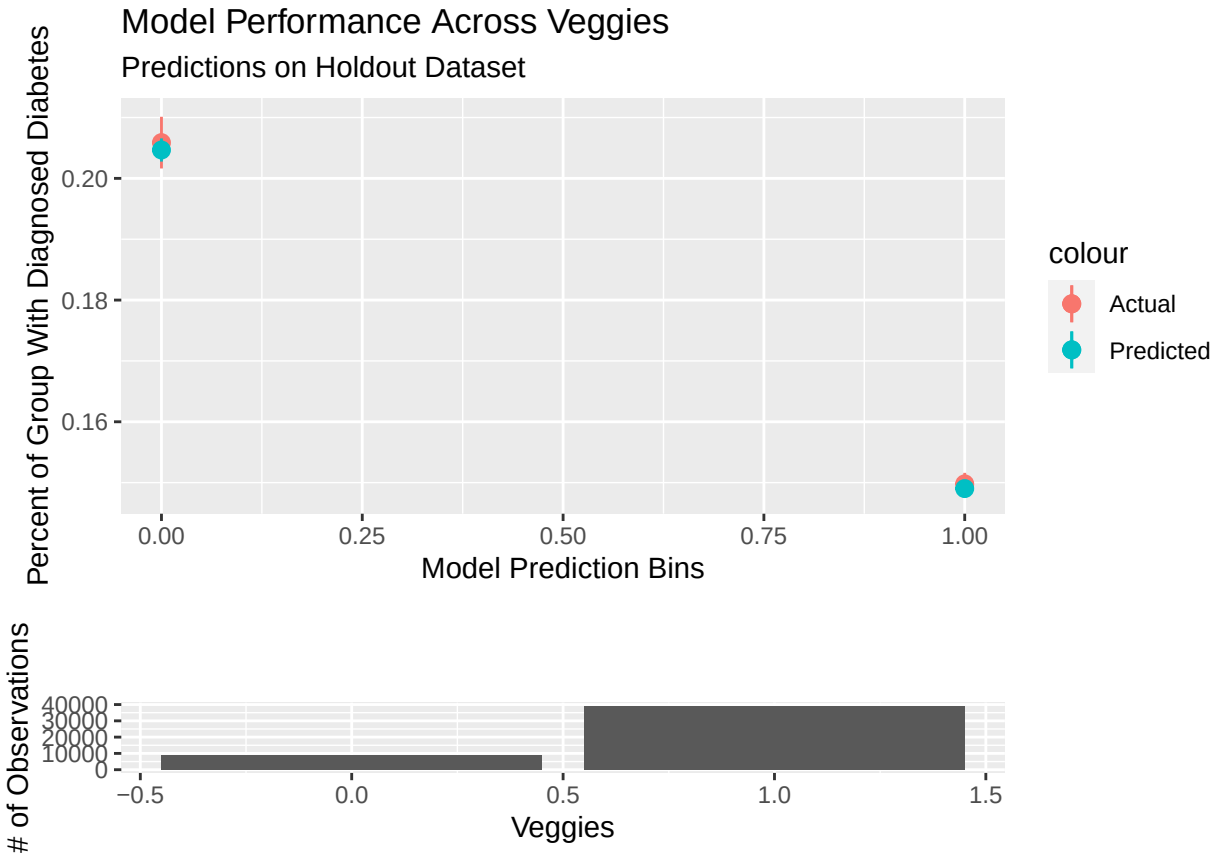
```
##
## [[7]]
```


Model Performance Across HvyAlcoholConsump
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.15
0.12
0.09
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
40000
30000
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
HvyAlcoholConsump

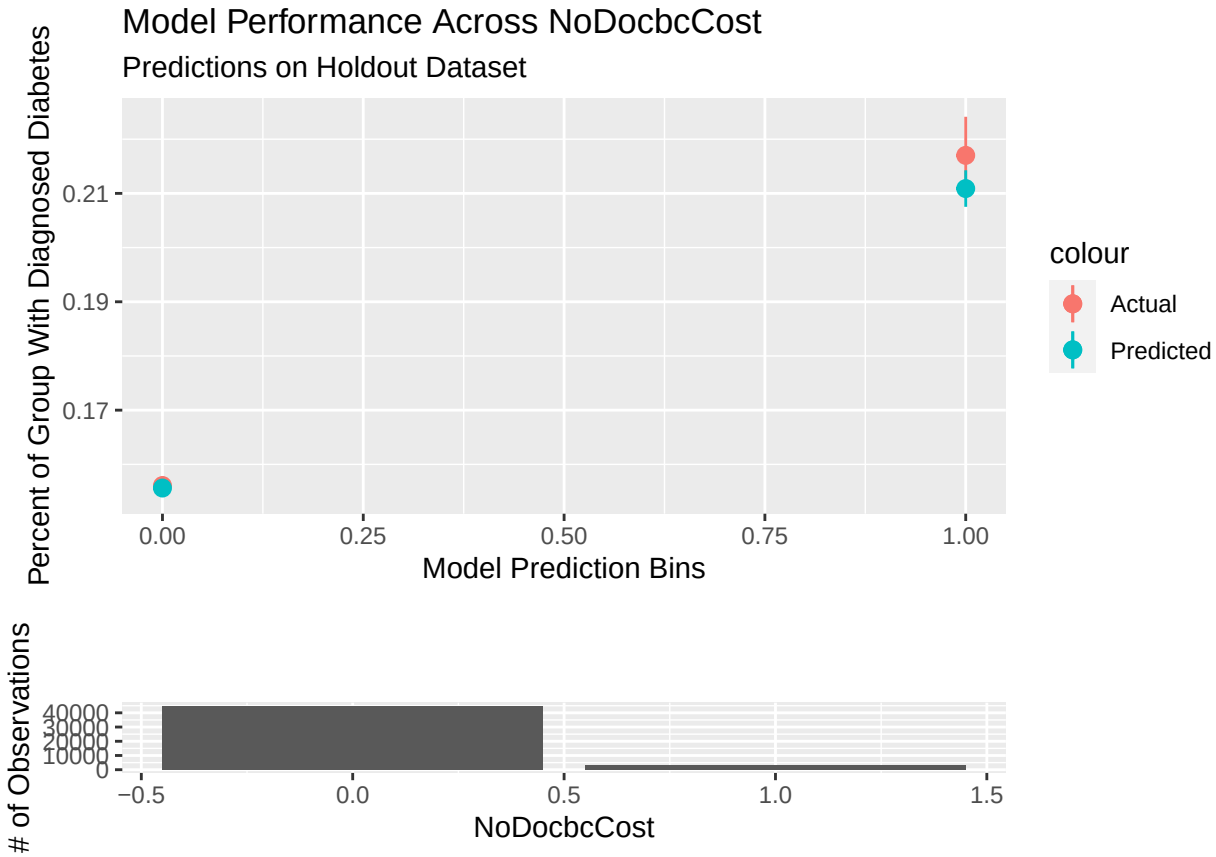
```
##
## [[8]]
```


Model Performance Across Fruits
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.18
0.17
0.16
0.15
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
30000
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
Fruits

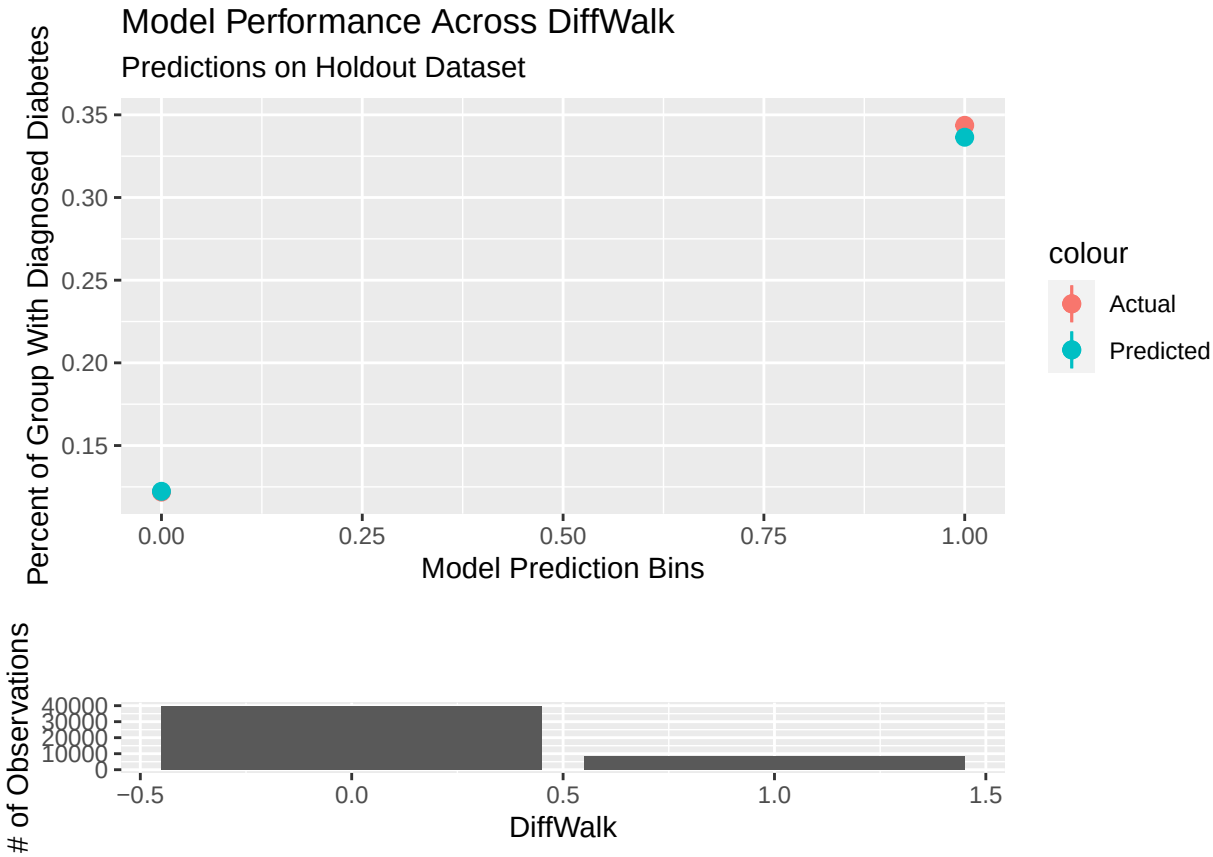
```
##
## [[9]]
```


Model Performance Across Veggies
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.20
0.18
0.16
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
40000
30000
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
Veggies

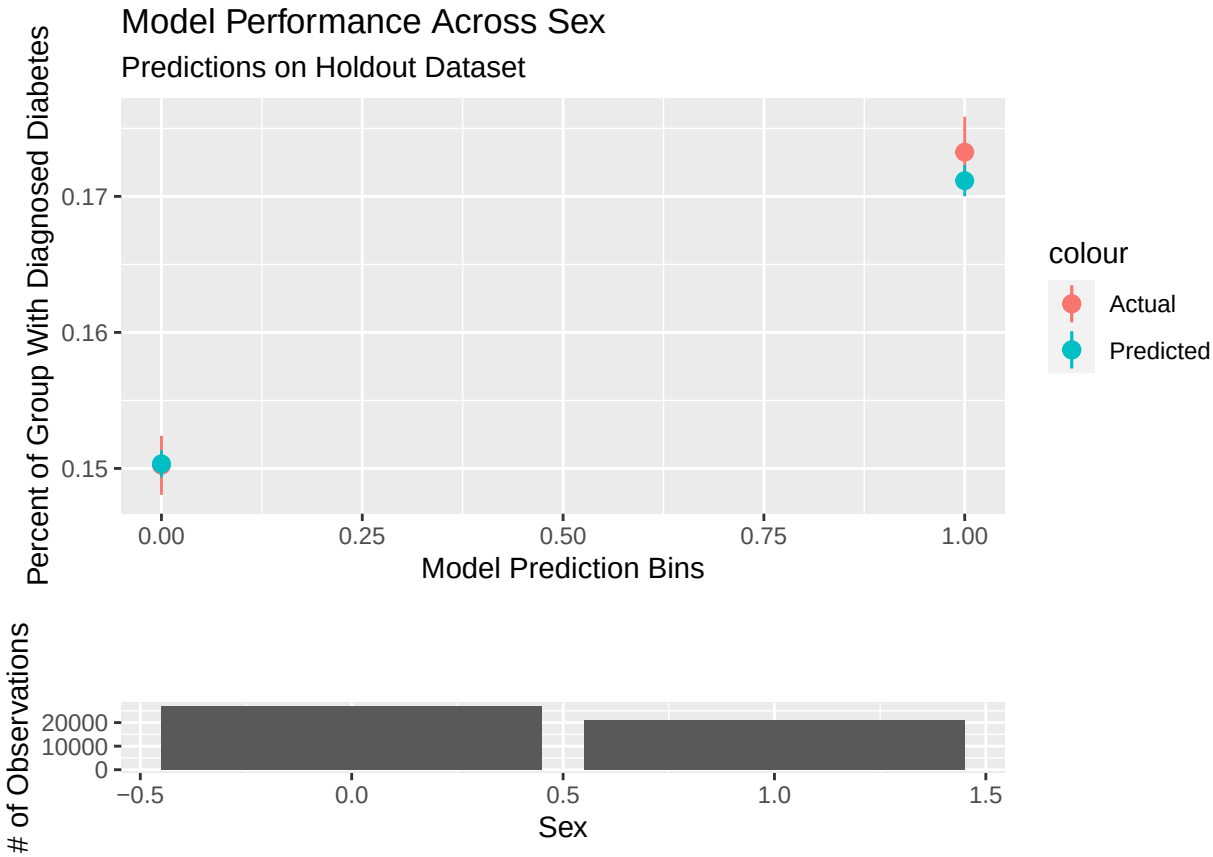
```
##
## [[10]]
```


Model Performance Across NoDocbcCost
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.21
0.19
0.17
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
40000
30000
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
NoDocbcCost

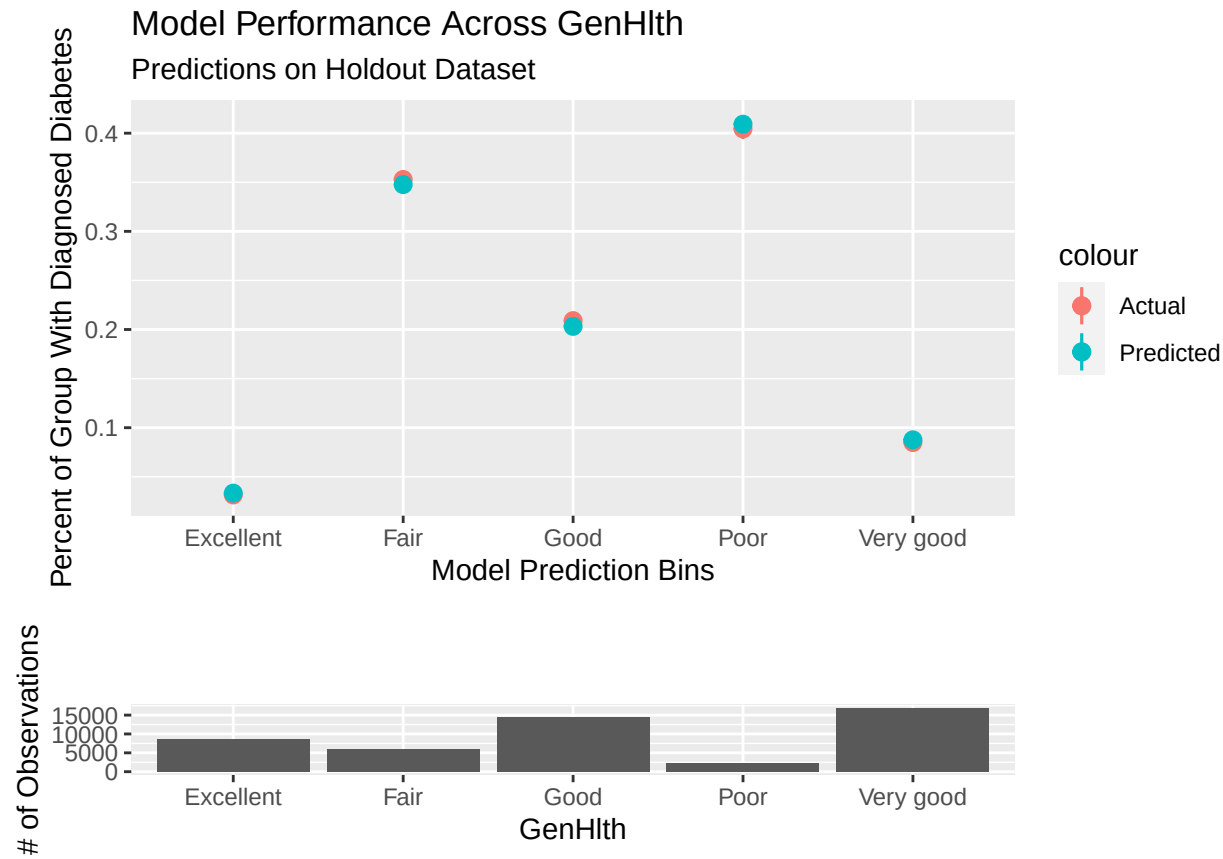
```
##
## [[11]]
```


Model Performance Across DiffWalk
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.35
0.30
0.25
0.20
0.15
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
40000
30000
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
DiffWalk

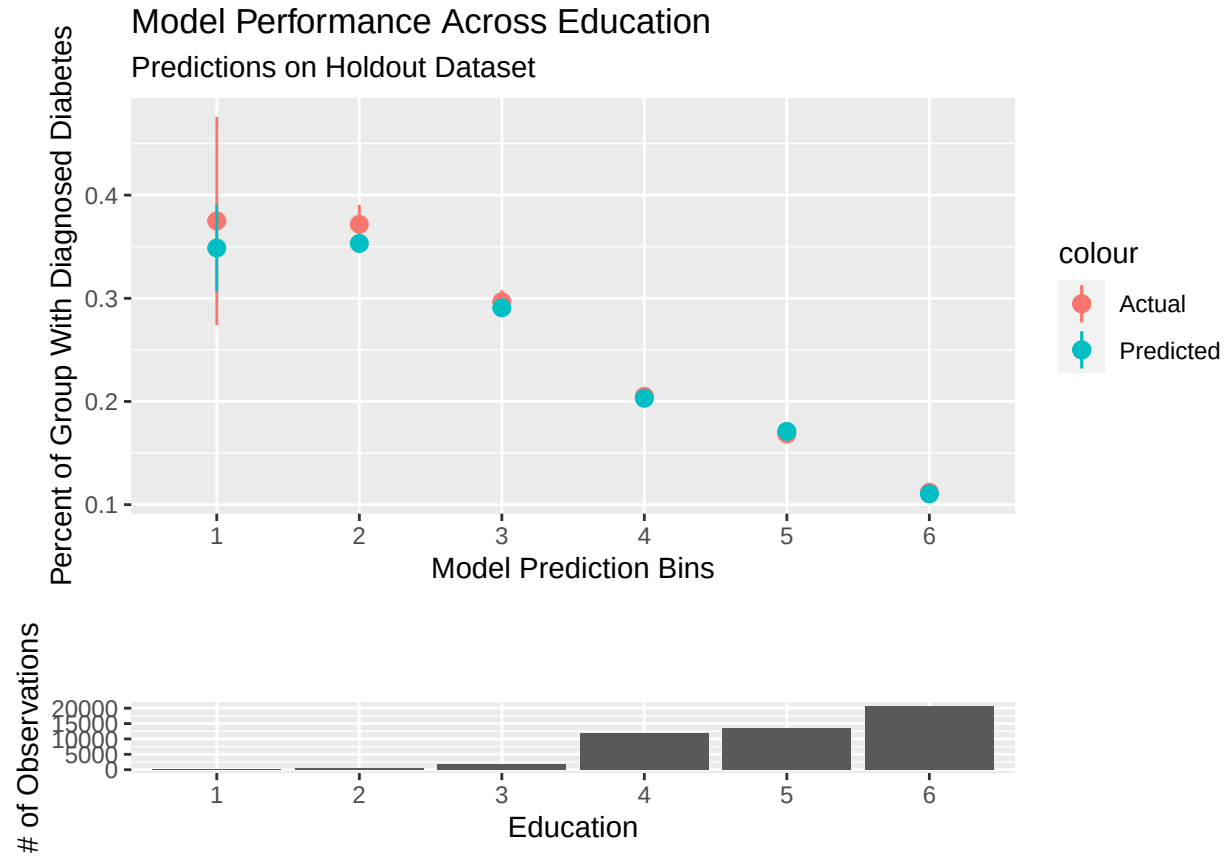
```
##
## [[12]]
```


Model Performance Across Sex
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.17
0.16
0.15
0.00
0.25
0.50
0.75
1.00
Model Prediction Bins
colour
Actual
Predicted
# of Observations
20000
10000
0
−0.5
0.0
0.5
1.0
1.5
Sex

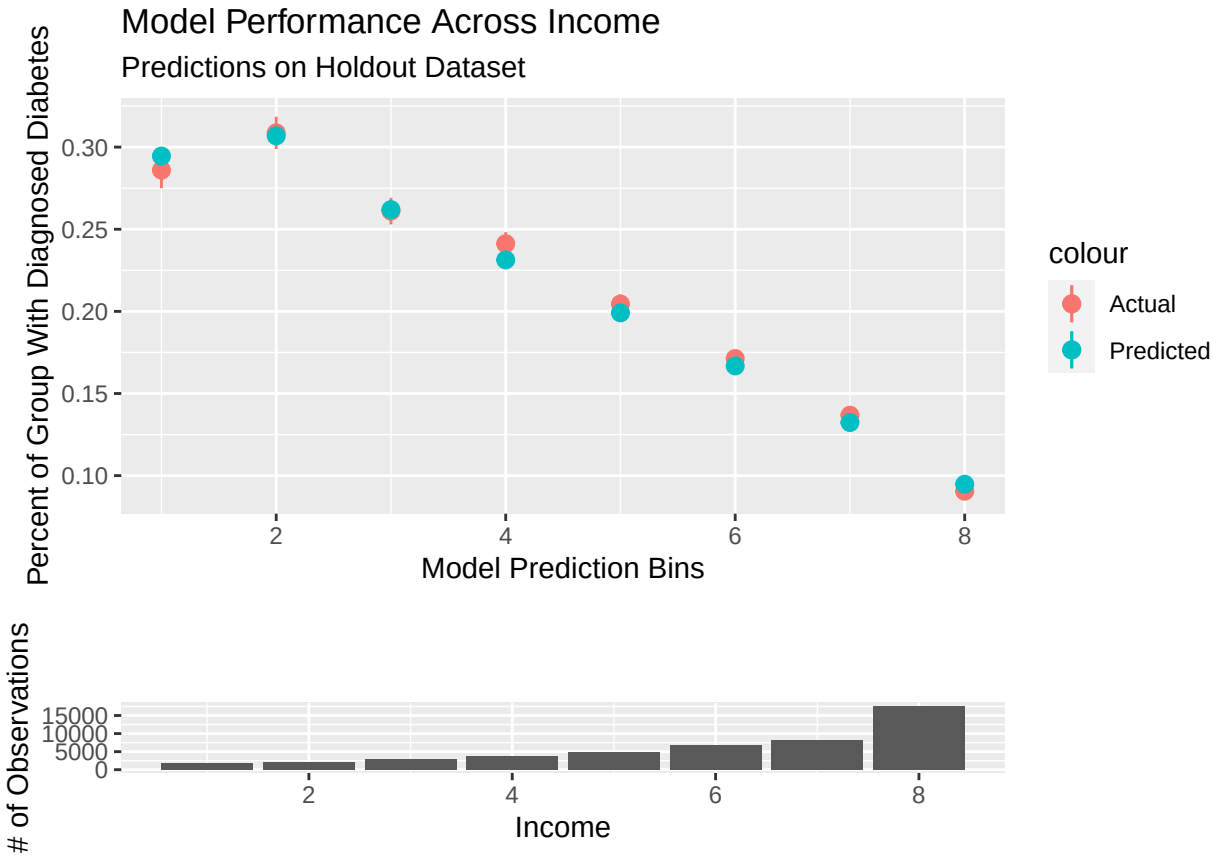
```
##
## [[13]]
```


Model Performance Across GenHlth
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.4
0.3
0.2
0.1
Excellent
Fair
Good
Poor
Very good
Model Prediction Bins
colour
Actual
Predicted
# of Observations
15000
10000
5000
0
Excellent
Fair
Good
Poor
Very good
GenHlth

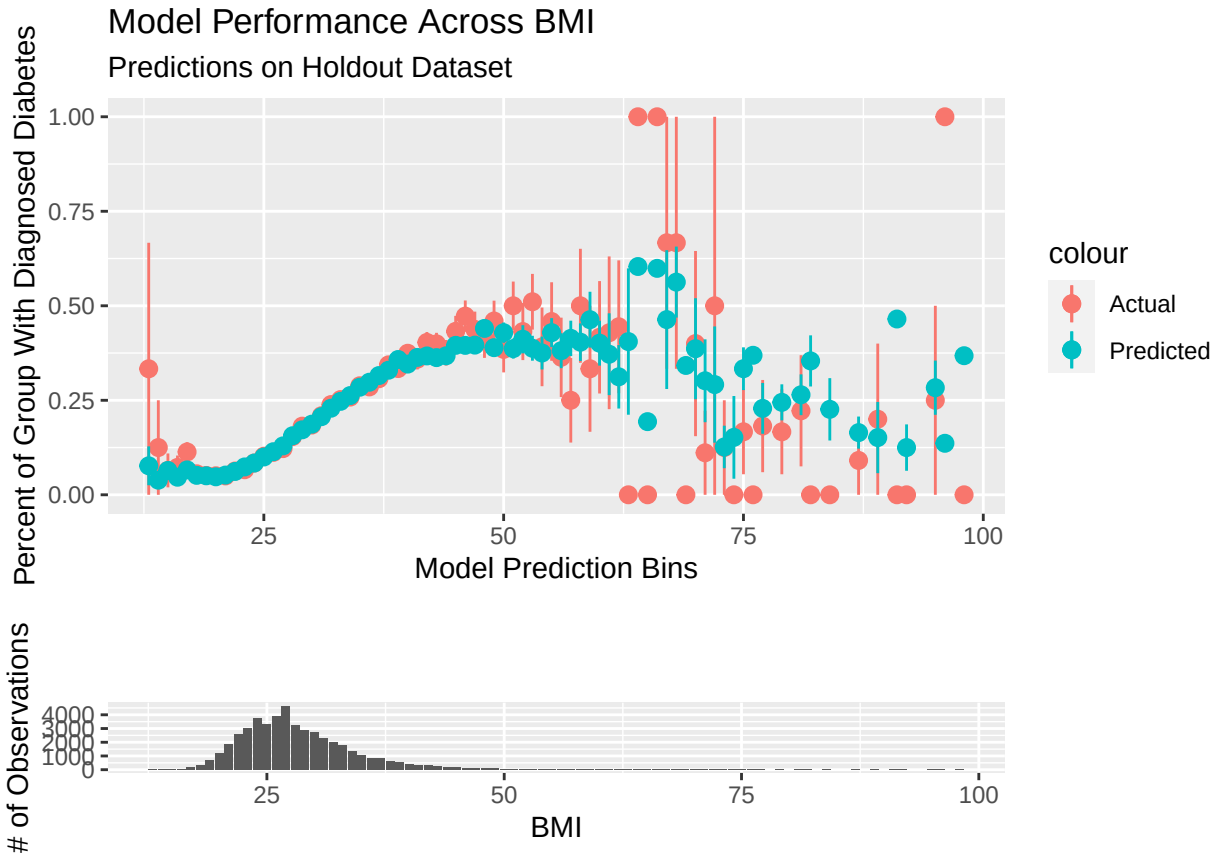
```
##
## [[14]]
```


Model Performance Across Education
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.4
0.3
0.2
0.1
1
2
3
4
5
6
Model Prediction Bins
colour
Actual
Predicted
# of Observations
20000
15000
10000
5000
0
1
2
3
4
5
6
Education

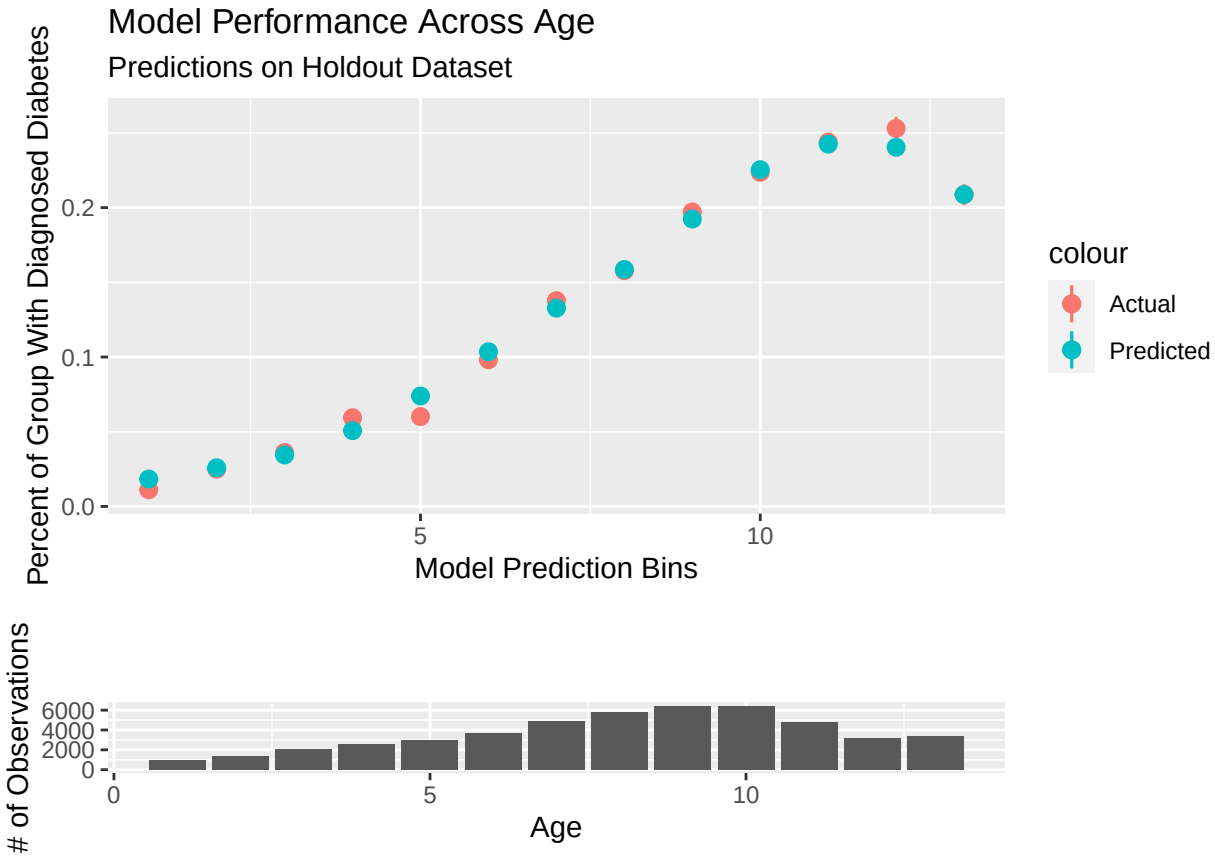
```
##
## [[15]]
```


Model Performance Across Income
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.30
0.25
0.20
0.15
0.10
2
4
6
8
Model Prediction Bins
colour
Actual
Predicted
# of Observations
15000
10000
5000
0
2
4
6
8
Income

```
##
## [[16]]
```


Model Performance Across BMI
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
1.00
0.75
0.50
0.25
0.00
25
50
75
100
Model Prediction Bins
colour
Actual
Predicted
# of Observations
4000
3000
2000
1000
0
25
50
75
100
BMI

```
##
## [[17]]
```


Model Performance Across Age
Predictions on Holdout Dataset
Percent of Group With Diagnosed Diabetes
0.0
0.1
0.2
5
10
Model Prediction Bins
colour
Actual
Predicted
# of Observations
0
2000
4000
6000
0
5
10
Age