Predicting Type II Diabetes: Analyzing Key Features with Logistic Regression

Group 1-
Rahul Singh, Lauren Blakeley,
Tania Amreen, and Aksheytha Chelikavada

Background and Motivation

Project Aim and Objective:
Identify and analyze key factors contributing to diabetes development, aiming for data-driven prevention strategies.

Prevention Potential:
Diabetes is a prevalent and costly chronic disease. Early prediction can support preventive care and resource planning.

Dataset Source:
"Diabetes Binary Health Indicators Dataset" from Kaggle, with 21 independent lifestyle, health, and demographic variables.

Approach:
We chose logistic regression on health and demographic factors for its interpretability and suitability for binary outcomes (diabetes: yes/no).

Key Questions

- **Question 1:**

  ***Impact of Independent Variables**:*

  Which independent variables have a significant impact on the development of Type II diabetes?

- **Question 2:**

  ***Predictive Power**:*

  Which combination of independent variables offers the best predictive power for determining whether an individual will or will not develop Type II diabetes?

Data Description

- **Dataset Overview:**

  Source: Behavioral Risk Factor Surveillance System (BRFSS) 2015, with 253,680 observations and 22 features.

  Target Variable: Diabetes_binary (1 = diabetes, 0 = no diabetes)

- **Features:**

  Includes lifestyle factors (e.g., smoking, physical activity), medical history (e.g., high blood pressure), and demographics.

- **Data Completeness**:

  No missing values; outliers observed in BMI, Physical Health, and Mental Health categories.

Methodology – Methodology Overview

- **Primary Model:**

  Logistic regression, selected for its interpretability with binary outcomes.

- **Stepwise Regression:**

  Used to select significant predictors. We chose to use both direction “forward selection” and “backward elimination” (dependent on the model) in our model building process.

- **Evaluation Metrics:**

  Confusion matrix: accuracy, sensitivityand AUC, deviance residuals for model performance.

Additional Techniques

- **Ridge & Lasso Regression Techniques:**

  Ridge and Lasso regression techniques are used to prevent overfitting by applying regularization to the model, with Ridge penalizing the sum of squared coefficients and Lasso performing both regularization and variable selection. We attempted both techniques and saw no improvement in our models.

- **Outcome**:

  Ridge didn’t improve accuracy, and Lasso reduced coefficients to zero. Stepwise regression was chosen for effective feature selection.

Methodology – Data Preprocessing

Handling Missing Values:

- Dataset was complete with no missing values.

Outlier Detection:

- Box plots helped identify and manage outliers in key variables like BMI and health-related measures.
  - The use of IQR
  - Removal of outliers

Feature Engineering:

- Grouped variables (e.g., Age and Physical Health) into categories for more effective analysis.
- Age <= 5 is Young, 5 < Age <=9 is Middle-Aged, Age > 9 is Older
- PhysHlth <= 5 is Low, 5 < PhysHlth <= 16 is Moderate, PhysHlth > 9 is High

Methodology- Model Building

- **Models Created (5):**

    1. Full logistic Regression Model
    2. Stepwise Regression for Feature Selection
    3. First-Order Logistic Regression Model
    4. Second-Order Logistic Regression Model with Interaction Terms and Squared Terms
    5. Interaction Model

- **VIF Values & Multicollinearity:**

    Looked at VIF values of predictors to look for multicollinearity. The highest VIF value observed was for *PhysHlthCategoryLow* at 2.31. We checked for multicollinearity to ensure our independent variables were not highly correlated as that could reduce the reliability of the inferences made from our models.

Methodology- VIF

Before

HighBP1	HighChol1	CholCheck1	BMI
1.1330	1.0727	1.0102	1.1353
Smoker1	Stroke1	HeartDiseaseorAttack1	PhysActivity1
1.0871	1.0734	1.1581	1.1400
Fruits1	Veggies1	HvyAlcoholConsump1	AnyHealthcare1
1.1125	1.1063	1.0125	1.0992
NoDocbcCost1	GenHlth2	GenHlth3	GenHlth4
1.1570	4.8137	6.3286	5.5805
GenHlth5	MentHlth	PhysHlth	DiffWalk1
3.9068	1.2850	1.8218	1.5056
Sex1	Age2	Age3	Age4
1.1169	2.7988	4.8552	8.5752
Age5	Age6	Age7	Age8
13.0010	19.1740	30.6560	38.6690
Age9	Age10	Age11	Age12
47.9000	51.9850	42.9130	30.8550
Age13	Education2	Education3	Education4
30.7560	23.7760	46.5870	178.5600
Education5	Education6	Income2	Income3
174.9200	183.8500	2.1414	2.3821
Income4	Income5	Income6	Income7
2.6562	2.9367	3.4128	3.5648
Income8			
4.7211			

After

CholCheck1	BMI	Smoker1
1.0098	1.0845	1.0715
PhysActivity1	Fruits1	Veggies1
1.1332	1.1064	1.1032
NoDocbcCost1	DiffWalk1	Sex1
1.1445	1.4821	1.1011
_levelLow Education	Education_levelMedium Education	Income_levelLow Income
1.0915	1.1543	1.2800
CategoryPoor Health	MentHlthCategoryLow	MentHlthCategoryModerate
1.7103	2.0275	1.7259

Results - Full Logistic Regression Model

Significant Predictors:

High blood pressure, cholesterol, BMI, stroke, heart disease, physical activity, age groups, gender, education, income, general health, and Heavy Alcohol Consumption were among these.

all features were kept in this model but there were quite a few predictors with p-values significant at .001. *HighBP1, HighChol1, CholCheck1, BMI, Stroke1, HeartDiseaseorAttack1, HvyAlcoholConsump1, DiffWalk1, Sex1, Age_groupOlder, Age_groupYoung, Education_levelLow_Education, Income_levelLowIncome, Income_levelMedium_Income, GenHealthCategoryGood_Health, and GenHealthCategoryPoor_Health* are those predictors.

Interpretation:

These predictors highlight the main health and demographic risk factors for diabetes.

```
$ConfusionMatrix
   resp.preds
        0     1
  0 41525   736
  1  5594   912

$Sensitivity
[1] 0.1401783

$Accuracy
[1] 0.8701991

$AUC
Area under the curve: 0.8224
```

Results - Stepwise Regression for Feature Selection

Significant Predictors:

Variables retained after stepwise selection include high blood pressure, cholesterol, physical activity, mental health levels, income levels, and health categories.

Features used in model:

HighBP1, HighChol1, CholCheck1, BMI, Stroke1, HeartDiseaseorAttack1, PhysActivity1, Veggies1, HvyAlcoholConsump1, AnyHealthcare1, DiffWalk1, Sex1, Age_groupOlder, Age_groupYoung, Education_levelLow Education, Education_levelMedium Education, Income_levelLow Income, Income_levelMedium Income, GenHealthCategoryGood_Health, GenHealthCategoryPoor_Health, MentHlthCategoryLow, MentHlthCategoryModerate

Interpretation:

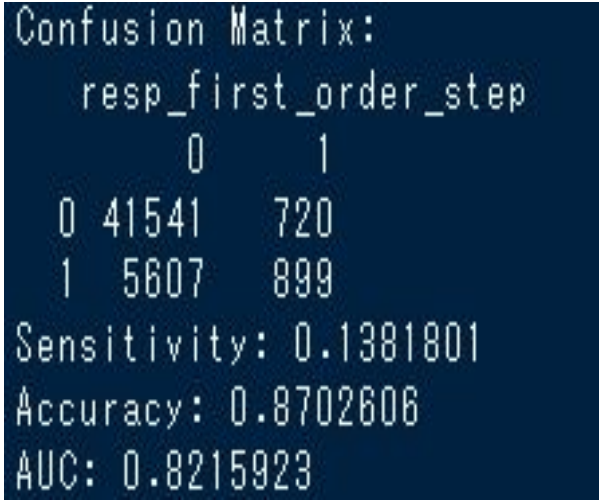
These predictors highlight the main health and demographic risk factors for diabetes.

```
$ConfusionMatrix
   resp.preds
        0     1
  0 41530   731
  1  5595   911

$Sensitivity
[1] 0.1400246

$Accuracy
[1] 0.8702811

$AUC
Area under the curve: 0.8223
```

Results - First Order Logistic Regression Model

Features used in model:

Age_groupOlder, Age_groupYoung, BMI, MentHlthCategoryLow, MentHlthCategoryModerate, HighBP1, HighChol1, CholCheck1, Stroke1, HeartDiseaseorAttack1, PhysActivity1, Fruits1, Veggies1, HvyAlcoholConsump1, NoDocbcCost1, GenHealthCategoryGood Health, GenHealthCategoryPoor Health, DiffWalk1, Sex1



Results - Second Order with Interaction Terms

- **Interaction Terms**:
  Age and Physical Health interactions provide a deeper understanding of how combined factors influence diabetes risk.
- **Model Complexity**:
  High number of interaction terms increased complexity but captured detailed relationships between variables.

```
Confusion Matrix:
   resp_second_order_step
         0     1
  0 41536   725
  1  5611   895
Sensitivity: 0.1375653
Accuracy: 0.8700761
AUC: 0.8223127
```

Features used in model:

Age_groupOlder, Age_groupYoung, BMI, MentHlthCategoryLow, MentHlthCategoryModerate, PhysHlthCategoryLow, PhysHlthCategoryModerate, Education_levelLow Education, Education_levelMedium Education, Income_levelLow Income, Income_levelMedium Income, HighBP1, HighChol1, CholCheck1, Stroke1, HeartDiseaseorAttack1, PhysActivity1, Veggies1, HvyAlcoholConsump1, AnyHealthcare1, GenHealthCategoryGood Health, GenHealthCategoryPoor Health, DiffWalk1, Sex1, Age_groupOlder:PhysHlthCategoryLow, Age_groupYoung:PhysHlthCategoryLow, Age_groupOlder:PhysHlthCategoryModerate, Age_groupYoung:PhysHlthCategoryModerate, BMI:PhysHlthCategoryLow, BMI:PhysHlthCategoryModerate

Results - Interaction Model

Features used in model:

There were 321 variables included in this interaction model after using forward selection to reduce the number of terms in the model. This large number of variables is derived from the inclusion of main effects and interaction terms between multiple categorical and continuous variables. This approach increases the model's complexity.

Attempts with Ridge and Lasso regression for feature selection were ineffective, as Lasso reduced all coefficients to zero.

Since it was very computationally expensive to use the "both" direction during Stepwise variable selection for this model due to the number of terms, the **forward selection** was chosen over the "both" direction, balancing feature selection and mode.

```
Confusion Matrix:
   resp_interaction_step
         0     1
  0 41700   561
  1  5717   789
Sensitivity: 0.1212727
Accuracy: 0.8712654
AUC: 0.824091
```

Interpretation of Results

- **Diabetes Predictors:**
  - There were many significant predictors of Type II diabetes in this model. Some were to be expected like BMI and high blood pressure. However, other independent variables such as income and heavy alcohol consumption having low and significant p-values is not necessarily common knowledge and worth exploring further.
- **Challenges in Prediction:**
  - The models have high Accuracy and AUC values but low sensitivity. High AUC means the model should be good at distinguishing between classes and the high accuracy means overall the models predict the data well but this could be because there are very few yes cases to which the model had low sensitivity.
- **Overall Performance**:
  - Strong accuracy but low sensitivity. There were many significant predictors.

Limitations

- **Data Limitations:**
  - Potential biases due to self-reported data.
- **Low Sensitivity:**
  - Indicates the model is poor at predicting true cases of type II diabetes.
- **Problem with overfitting:**
  - Due to a small number of yes cases for type II diabetes in the dataset, our model is likely overfit to the training data samples for yes cases- capturing too much noise and variance within this data and reducing its predictive capabilities.

- **Assumes linearity between the dependent and independent variables:**
  - Our variables do not have a linear relationship with having or not having type II diabetes (0 or 1)

Future Improvements

- **Advanced Models:**
  - Exploring more complex models like random forests or neural networks.

- **Expand Dataset:**
  - Use more recent or additional datasets for broader applicability.

- **Include Additional Variables:**
  - Identify and incorporate new health-related variables for improved accuracy.

Conclusion

- **Summary:**
  - Logistic regression and Lasso allowed us to create an interpretable model with good accuracy.

- **Impact:**
  - Model can help identify high-risk individuals for early intervention.

- **Final Thought:**
  - Improving sensitivity and incorporating more data could further enhance model utility.

Questions

Thank You!