NATIONAL UNIVERSITY OF SINGAPORE

FACULTY OF SCIENCE

AY 2021/2022, SEMESTER 2
ST4248: Statistical Learning II

Group Project Final Report

Title:
Predicting Diabetic status amongst respondents

Group B2:

CHAN WEN YONG (A0201868B),
CHEW BANGYAO PAUL (A0204816J),
LIM XI CHEN TERRY (A0199513W),
LOW HON ZHENG (A0204803R)

Summary:

In this report, we seek to accurately predict if an individual is diabetic based on their health profile through exploring different types of models. Additionally, we also seek to identify the leading contributors to the development of diabetes. This will be useful as an early warning sign and can serve as valuable information on habits that individuals should take up or modify in order to reduce their risk of developing diabetes. We will be sharing more on data collection, preprocessing, models, limitations and finally the conclusion gained from our experimentation.

Table of contents

Introduction

Diabetes is a chronic disease which can lead to blindness, kidney failures and even lower limb amputations. In 2019, the World Health Organisation reported that diabetes was the direct cause of 1.5 million deaths with premature mortality rates (deaths before the age of 70) increasing by 5% between 2000 and 2016.[1] It is also described as a 'silent' disease in its early stages, where patients can feel perfectly well until complications arise. However, a late diagnosis can result in serious and irreversible complications. In the past three decades, the prevalence of diabetes has risen dramatically across countries of all income levels.[2]

Thus, this creates the need for an early detection system to identify potential diabetic patients, so that they can obtain timely treatment.

Description of data

There are 3 different types of diabetes. Type 1 diabetes is characterised by deficient insulin production and requires daily usage of insulin. Neither its cause nor the means to prevent it are known. Type 2 diabetes results from the body's ineffective use of insulin. This type of diabetes generally occurs over the course of one's life as a result of lifestyle factors. Lastly, we have gestational diabetes which occurs during pregnancy, where women have blood glucose values above normal levels but below those of Type 1 and Type 2 diabetic patients during their course of pregnancy.[1]

The Behavioural Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that has been conducted annually since 1984 by the CDC (Centres for Disease Control and Prevention) in America. Responses were collected from over 400,000 Americans on their daily behaviours, as well as their chronic health conditions.

For this project, the focus will be on the latest published data from 2020, which contains 279 questions and responses from 401,958 different individuals.

Taking a deeper look at the `IS_DIABETIC` column from the dataset, we noticed that we could not accurately and conclusively determine which of the respondents were suffering from Type 1, Type 2 or gestational diabetes. As such, we made the assumption that all diabetic respondents in this dataset refer to Type 2 diabetes.

Problem Statement

We are interested to find out how an individual's profile would determine whether they are likely to be diabetic or not. Through this, we hope that more diabetic patients can be identified at the early stages of their disease before their conditions worsen, increasing their chances of survival.

We aim to do this through a 3-step approach:

1. Predicting if a respondent is diabetic using demographics and health-related data.
2. Identifying the most important features in the dataset that lead to a higher risk of diabetes.
3. Using the subset of important features to accurately predict whether an individual has diabetes.

Data Exploration and Cleaning

Feature Extraction

After analysing the 279 features based on the codebook, we selected IS_DIABETIC as our response as well as a subset of features that have possible relationships to a respondent developing diabetes. We are able to categorise these 21 features into the following 3 main groups:

1. **Demographics** (BMI (BMI), Sex (SEX), Age (AGEGRP), Marital Status (MARITALGRP), Income Group (INCOMEGRP), Race (RACE), Unable to see doctor due to cost (HAS_MONEYPROB))
2. **General Health** (Days with poor Physical Health (NUM_POORPHYHLTH), Days with poor Mental Health (NUM_POORMENTHLTH), Has Stroke (HAS_STROKE), Has Heart Disease (HAS_HD), Has Healthcare Plan (HAS_HLTHPLAN), Difficulty Walking (HAS_DIFFWALK), Days since last checkup (CHECKUP), General Health Level (GENHLTH_LVL), Days unable to perform usual activities due to poor health (NUM_POORHLTH))
3. **Behavioural patterns** (Smoking Frequency (SMOKER_TYPE), Did Physical Activity in past month (HAS_PHYACT), Number of Alcoholic drinks drank per week (NUM_DRINKSPERWK), Frequency of use of E-Cigarette (HAS_ECIG), Hours of Sleep in a day (NUM_SLEEP))

The features that are continuous include BMI, NUM_DRINKSPERWK, NUM_SLEEP, NUM_POORMENTHLTH, NUM_POORPHYHLTH, NUM_POORHLTH, while the rest are categorical.

Data Cleaning

The dataset was discovered to possess the following deficiencies: missing values; ambiguous values in certain features; outliers in the continuous features; correlated variables. In this following section, we will discuss the steps we took to address these flaws within the dataset.

- Missing Values

We noticed that 2 features, HAS_ECIG and NUM_POORHLTH have a large number of missing values as compared to the rest of the features. These 2 features had 87.4% and 49.8% of values missing respectively compared to less than 10.3% across the rest of the variables in the dataset. Instead of simply removing these missing values (which may potentially be the bulk of the data), we investigated the reason behind these missing entries. Examining the codebook, a possible reason as to why these questions were left empty by a large proportion of respondents was because they were told to skip these questions due to their responses to preceding questions.

Take for example, the case for NUM_POORHLTH. Respondents were asked to leave this question blank if in the past 30 days there were not any days where their physical and mental health were not good. Thus, for this feature, we checked for respondents who left NUM_POORHLTH empty while replying “None” in NUM_POORPHYSHLTH and NUM_POORMENTHLTH. For such respondents, we converted the blank values in NUM_POORHLTH to “None”. For HAS_ECIG, based on the preceding questions asked, we noticed that

most of the blank entries signified that the respondent has never used an e-cigarette or other electronic vaping products in their life. Hence, we created a new level 0 to represent these respondents.

As mentioned earlier, since the percentage of blank entries in the other variables was not as prevalent, we removed all rows which possessed missing values across the variables. This resulted in a dataset of 242,301 rows and 22 variables.

- Ambiguous responses in questions

From further exploration of the data, we also noticed that in a number of the questions posed, there were options like “Don’t know/Not Sure” or “Refused” for respondents to select. As we believed that these entries introduced ambiguity and affected possible interpretation and recommendations of any model built, we decided to remove survey data of respondents who provided such responses. This reduced the dataset to 187,823 rows of entries.

- Outliers

We then proceeded to detect if there were outliers present in any of the features. We skipped this step for the categorical features since there is no way to detect outliers for them. For the remaining 6 continuous features, we identified potential outliers from each of these features by studying observations that lay outside of the whiskers of the boxplot. After analysing the boxplots together with contextual knowledge, we found that the only variable that possessed unreasonable values was the `NUM_DRINKSPERWK` feature displaying the number of alcoholic drinks drank per week. The values for this feature were obtained by multiplying the average number of drinks drank per drinking occasion * number of drink occasions per day * 7. One drink is equivalent to a 12-ounce beer, a 5-ounce glass of wine or a drink with one shot of liquor. The boxplot for this feature can be seen below in Figure 1, with the upper whisker values being 747.

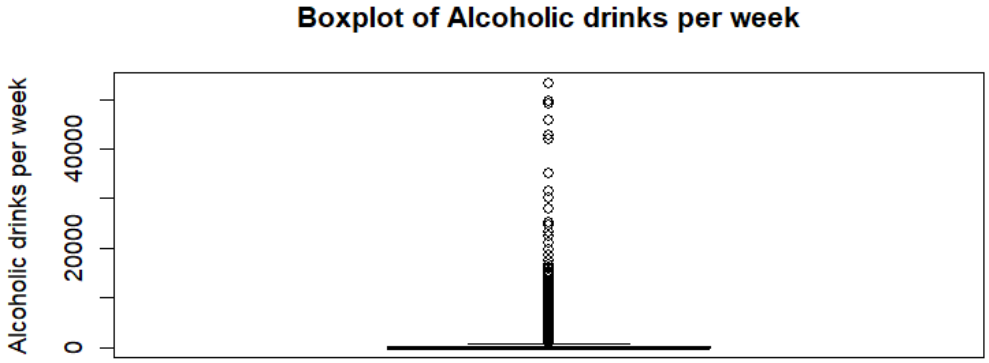

Figure 1: Boxplot returned for the Number of Alcoholic drinks drank per week

As we agreed that this value of 747 is a fair upper bound, our group decided to remove the 21,456 observations that had values larger than this. This reduced the dataset to 166,367 rows of entries.

- Correlated variables

Correlation between the variables was also checked to reduce multicollinearity in our dataset. As there is a mix of categorical and continuous variables in our dataset, we used various methods to calculate the correlation between each possible pair of variables, and the results can be seen in Table 1 below.

Comparison	Method	Analysis	Result
Continuous variables	Pearson Correlation Coefficient	Moderate correlation found between: 1. `NUM_POORHLTH` and `NUM_POORPHYHLTH` (0.611) 2. `NUMPOORHLTH` and `NUM_POORMENTHLTH` (0.451)	Remove `NUMPOORHLTH`
Binary Categorical variables	Cramér's V	All values < 0.5 and are close to 0	No correlation found
Non-Binary Categorical variables	Phi Coefficient	All values are either < 0.5 or > -0.5 and are close to 0	
Continuous & Binary Categorical variables	Point-Biserial Correlation Coefficient	All values are either < 0.5 or > -0.5 and are close to 0	

Table 1: Correlation analysis of features in the dataset

With the removal of `NUM_POORHLTH`, this resulted in a final data set of 21 variables and 166,367 rows of entries.

Data Exploration

Let us share more about some of the exploratory data analysis that was conducted on the variables in the dataset. These are some variables that we found are likely to increase the risk of developing diabetes.

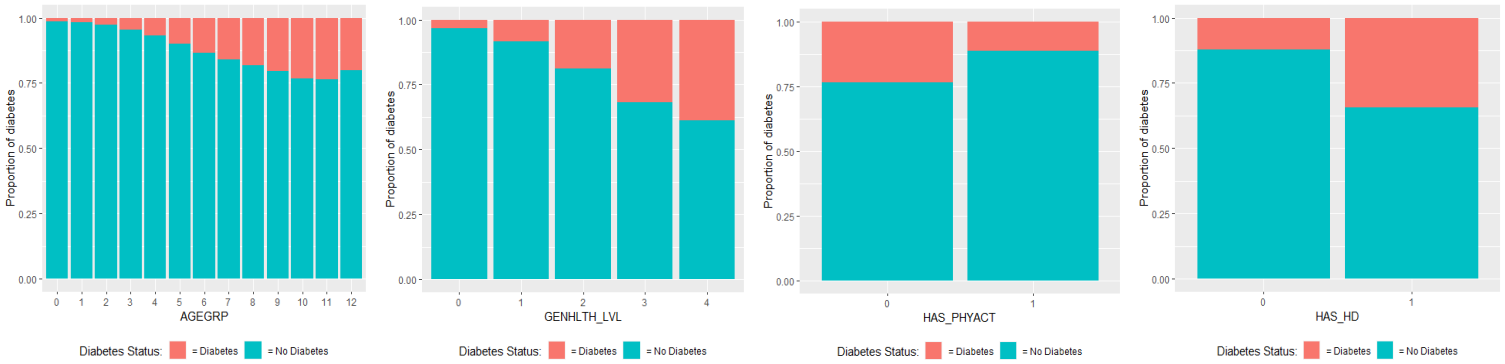

Figure 2: Plot of proportion of diabetes in various categorical variables

From these plots, there is a clear association between diabetes and each of these variables such that the rate of diabetes differs in each category of the variable.

Taking a deeper look at the `IS_DIABETIC` response, we noticed a clear imbalance between the number of respondents who have diabetes and the number of respondents who do not have diabetes, with only 14.1% of the dataset comprising respondents with diabetes. This imbalance in the 2 classes in our dataset may cause a problem in our modelling process and will be dealt with in the later sections.

Models

After carrying out data exploration and data cleaning, we proceeded to build our models. We decided to carry out a 70-30 train test split on the 166,367 rows, leaving 116,458 rows of training data and 49,909 rows of testing data. Training data is used to fit our models, while the test data is used to evaluate the model's performance on new unseen data.

The following metrics are used to evaluate the models' performance during fitting: **accuracy, precision, recall, F1 score,** and **AUC**. Due to the nature of this project, having a low recall or a low precision is detrimental for the model since it will either mean that the model will incorrectly predict many respondents with diabetes as not having diabetes, failing its purpose as an early detection system or incorrectly predict many respondents without diabetes as having diabetes, putting a heavy strain on medical institutions. Since our primary aim is to devise an early detection system, we will place a heavier emphasis on recall as compared to precision.

Baseline Model (Logistic Regression model)

We started off with the logistic regression model. This was chosen to be our baseline model because it is a simple yet effective classification algorithm commonly used for binary class classification due to its low cost and interpretability. We set `IS_DIABETIC` to be the response while the rest of the variables from our final dataset were set as the predictors. No additional transformation was done on the training set (such as downsampling and scaling) since this model will only serve as a baseline.

The logistic regression model was built using the glm() function from the stats library, while setting the family to be binomial to simulate a binary classification model. Setting a threshold of 0.5 to determine if a data point will be assigned to class 0 or class 1, the following confusion matrix was obtained when used to predict the test set.

	truth	
predict	0	1
0	41959	6007
1	893	1050

Figure 3: Confusion matrix for the Baseline model

Accuracy	Precision	Recall	F1 score	AUC
0.862	0.540	0.149	0.233	0.564

Table 2: Baseline model's test set performance

From Table 2, it can be seen that even though the model had a fairly high accuracy, it performed extremely poorly in the recall metric. This meant that the model does poorly in correctly identifying respondents with diabetes as seen from the large number of False Negatives in Figure 3. This is especially dangerous for models that are meant to be used in the medical industry, as this can result in patients who actually have diabetes being incorrectly predicted as not having diabetes by the model.

We hypothesised that the poor results obtained were due to the imbalanced nature of the dataset. Hence, from our analysis, we decided to carry out downsampling using the downSample() function from the caret library on the training data to ensure an equal distribution of both classes. This was only implemented on the training data to prevent the introduction of bias to the distribution of the test set. This results in both classes having 16,469 rows of data each in the downsampled training data.

L1 Regularised Logistic Regression model

Another hypothesis was that the baseline model built may have performed poorly in some of the metrics because of the presence of noisy features. Hence, we attempted a regularised variation of the logistic regression model to experiment if the addition of regularisation can effectively reduce the feature space and return better results. We decided to employ the lasso regularisation over the ridge regularisation because of lasso's greater interpretability from its variable selection capability. The lasso coefficients minimise the following objective function.

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Figure 4: Objective function of Lasso Regression model

Before fitting this model, one-hot encoding is to be done on the categorical features. The glmnet() function from the glmnet library was then used for the fitting of the model. The parameters alpha = 1 and family = "binomial" were set to specify that it is a 2-class lasso regularised logistic regression model. For the tuning parameter λ, we retained the sequence chosen by glmnet as it has a higher chance of attaining convergence. As glmnet() standardised variables by default, no scaling of the dataset was needed. Cross-validation was used to determine the optimal value for the λ hyperparameter and the cross-validation error returned for the different values of λ can be seen in Figure 5 below.

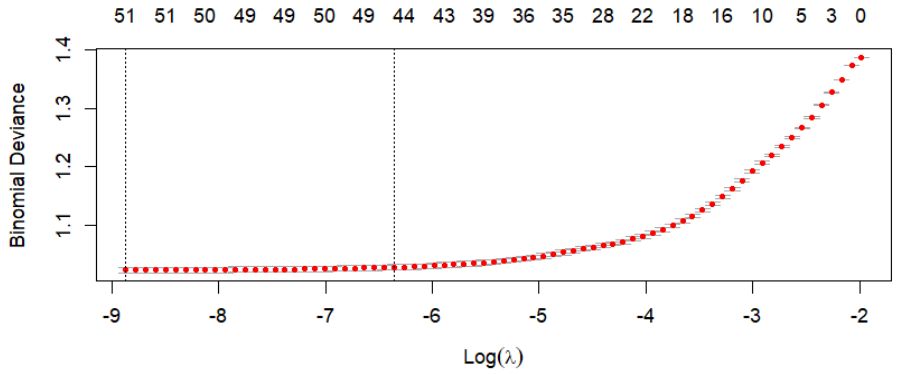

Figure 5: Cross-validation error for various log(λ) values

The optimal λ value returned was λ = 0.000141, which corresponds to Log(λ) = -8.87. Using this value of λ that minimised cross-validation error, we fitted the regularised logistic regression model on the downsampled one-hot encoded training data and evaluated its performance on the test set. The resulting model suggested removing 1 variable, `HAS_ECIG1`. Setting the threshold of the logistic regression model to 0.5, the predictions made by this model on the test set data can be seen in Figure 6 below while its performance across the 5 metrics can be seen in Table 3.

```
         truth
predict     0     1
      0 31148  1728
      1 11704  5329
```

Figure 6: Confusion matrix for the L1 Regularised Logistic Regression model

Accuracy	Precision	Recall	F1 score	AUC
0.731	0.313	0.755	0.442	0.741

Table 3: L1 Regularised Logistic Regression model's test set performance

Evaluating the model's performance across the 5 metrics, we note that even though recall has significantly improved, precision has decreased drastically. This means that the current model will often predict respondents without diabetes as having diabetes. This is not ideal as it will increase the workload of medical institutions.

eXtreme Gradient Boosting model (XGBoost)

Given our low precision score for the Regularised Logistic Regression model, our next hypothesis is that there might be non-linearity present in the data. Thus, we proposed to use the XGBoost model to better explain this non-linearity.

XGBoost is a tree-based ensemble machine learning algorithm. Due to how it improvises the Gradient Boosting framework by introducing accurate approximation algorithms, it often has higher predictive power and performance.[3] For these models, trees are grown in a sequential manner, which converts weak learners into strong learners by adding weights to the weak learners and reducing the weights of the strong learners. This results in each tree learning and boosting from the previous tree grown. XGBoost also often performs well in Kaggle competitions, making it a popular classification model.

For our model fitting process, we first employed randomised search to find the best set of parameters that has the lowest validation error within the large parameter space. Setting a threshold of 0.5 to determine if a data point will be assigned to class 0 or class 1, Figure 7 was obtained when predicting the test set.

	truth	
predict	0	1
0	32046	1828
1	10806	5229

Figure 7: Confusion matrix for the XGBoost model

Accuracy	Precision	Recall	F1 score	AUC
0.747	0.326	0.741	0.452	0.744

Table 4: XGBoost model's test set performance

Evaluating the scores from the 5 metrics in Table 4, it can be seen that even though the model performs well in recall, its precision remains low at 0.326. Therefore, a more sophisticated model might be needed to explain the variability in the dataset.

Neural Network

We experimented using a more complex model for better results, such as a Neural Network model. We chose Neural Networks because of its high tolerance to noise as well as its ability to classify patterns on data that they have not been trained on.[4]

To help speed up convergence and stabilise the learning process, we scaled the numerical features and one-hot encoded the categorical variables. Thereafter, to avoid the bias caused by an imbalanced dataset, we downsampled the training data before feeding it into a 3 hidden layer neural network consisting of 6,591 parameters.

Structure of our Neural Network model: 1. Input layer of 53 units; 2. First hidden layer of 100 units; 3. Second hidden layer of 50 units; 4. Third hidden layer of 20 units; 5. Output layer of 1 unit. Each hidden layer utilises the ReLU activation function and has a dropout

regularisation of p = 0.2. These hyperparameters of the Neural Network were decided through an iterative process of experimentation in an attempt to increase the validation accuracy of the model. The model was fitted using the R torch package and trained on 15 epochs of training data.

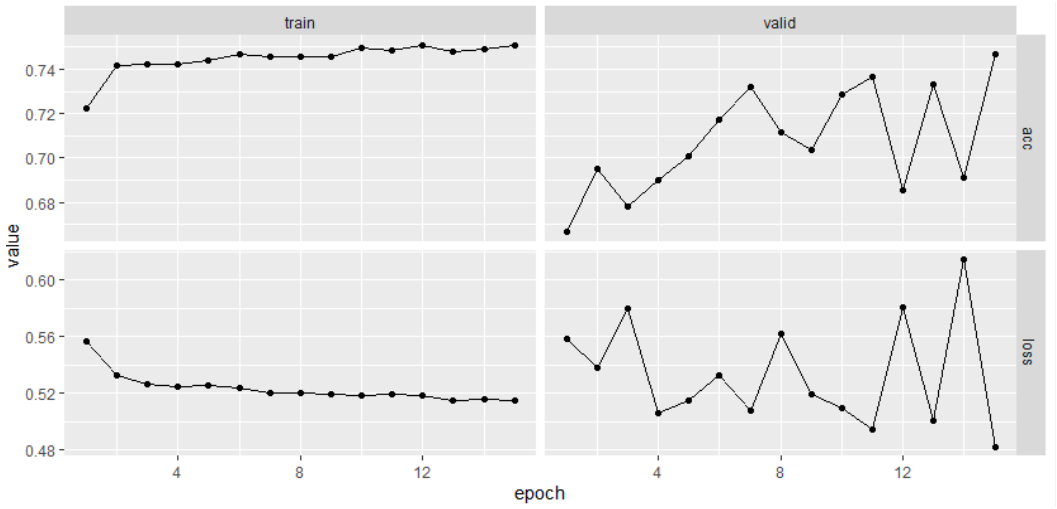

Figure 8: Plot of the Neural Network model

Figure 8 shows the performance of the model across the different epochs during the fitting process. Our initial try of 30 epochs shows that the validation accuracy decreased significantly after epoch 15, signifying that overfitting has occurred. Therefore, we decided to train the model for only 15 epochs. Setting a threshold of 0.5 to determine if a data point will be assigned to class 0 or class 1, Figure 9 was obtained when predicting the test set.

	truth	
predict	0	1
0	32093	1876
1	10759	5181

Figure 9: Confusion matrix for the Neural Network model

Accuracy	Precision	Recall	F1 score	AUC
0.747	0.325	0.734	0.451	0.742

Table 5: Neural Network model's test set performance

The results obtained are similar to the one obtained by the XGBoost and L1 Regularised Logistic Regression model. However, we also note that neural networks take a significantly longer time for training, where approximately 6 - 7 minutes were needed to train the model for 15 epochs.

Comparing models

	Logistic Regression (Baseline)	L1 Regularised Logistic Regression	XGBoost	Neural Network
Accuracy	0.862	0.731	0.747	0.747
Precision	0.540	0.313	0.326	0.325

Recall	0.149	0.755	0.741	0.734
F1 score	0.233	0.442	0.452	0.451
AUC	0.564	0.741	0.744	0.742

Table 6: Performance of the 4 models built

From Table 6, we can see the performance of the 4 models built across the 5 selected metrics. Comparing the performances of the different models, the L1 Regularised Logistic Regression model was chosen as the best model because of its high recall score and its ease of interpretability compared to the other models. This also serves as an indication that using a subset of features may be beneficial for improving model performance due to the potential presence of noisy features. This coincidentally ties in nicely with the next step of our 3-step approach, which is to identify the most important features in our dataset.

Variable Selection

Subset of features selected

To determine a subset of features, the stepAIC() function in the MASS package was used to perform forward and backward stepwise regression. Since the objective is to select a subset of features that are useful for inference, BIC was chosen as the criterion in determining the best model, as BIC tends to favour smaller subsets and would tend to select the correct variables. From the original 20 predictors, we obtain a subset of 11 predictors which are:

1. **Demographics** (BMI (`BMI`), Sex (`SEX`), Age (`AGEGRP`), Race (`RACE`), Unable to see doctor due to cost (`HAS_MONEYPROB`))
2. **General Health** (Has Stroke (`HAS_STROKE`), Has Heart Disease (`HAS_HD`), Days since last checkup (`CHECKUP`), General Health Level (`GENHLTH_LVL`))
3. **Behavioural patterns** (Did Physical Activity in past month (`HAS_PHYACT`), Number of Alcoholic drinks drank per week (`NUM_DRINKSPERWK`))

Model built on subset of features selected

Using the best model from Comparing Models and the subset of features, we evaluated its performance on the test set as shown in Figure 10 and Table 7.

```
           truth
prediction     0     1
         0 30574  1555
         1 12278  5502
```

Figure 10: Confusion matrix for logistic regression with a subset of features

	Accuracy	Precision	Recall	F1 score	AUC
Logistic Regression model with subset of features	0.723	0.309	0.780	0.443	0.747
L1 Regularised Logistic Regression model	0.731	0.313	0.755	0.442	0.741

Table 7: Logistic regression using subset of features's vs current best model test set performance

We note that this model performs as well as the other models built earlier and also uses fewer features. This suggests that the subset of features is sufficient in explaining the variability in the response and adding the remaining variables would only give a marginal increase in performance. Furthermore, the interpretability of a logistic regression model allows us to quantify the effects of the variables on the risk of diabetes.

Limitations

Key features missing from the dataset

From the results of the models, we notice that their performances are roughly similar, even with the tuning of hyperparameters and regularisation. Additionally, all the models either performed well in precision or recall and poorly for the other. This could mean that other significant variables might be needed to better explain the variance in the diabetic status of respondents. Examining the questions asked in our dataset, it is found that variables such as blood pressure, cholesterol and diet were not available in the dataset. These are key factors that can affect the respondent's risk of diabetes. Therefore, adding these variables may allow the models to better differentiate between the two classes and perhaps build one that does well in both recall and precision.

Missing details about the response variable

In addition, having more information with regards to the type of diabetes present amongst respondents may help to improve the performance of the model. This will allow us to only use the information from actual Type 2 diabetic patients who have developed diabetes due to their lifestyle habits. Thus, we can better pinpoint actionable insights into lifestyle habits that can be worked on to reduce the risk of developing diabetes.

Conclusion

In this project, we have explored various models and compared their performance in predicting diabetic patients. While the performance leaves more to be desired on a whole, the huge improvement to recall over the baseline model means that our models are able to do well in identifying patients with diabetes. Thus, it is able to function as an early detection model to allow for timely intervention and reduce the mortality rate from diabetes. Additionally, we have identified a subset of predictors that are useful in predicting diabetes using stepwise regression models. This subset of predictors makes it easier for individuals and doctors to quickly assess patients' risk of diabetes and understand which predictors are the ones that are increasing the risk of developing diabetes. Also, this subset of predictors can serve as a general guideline for healthy living to decrease the risk of diabetes, which will ultimately reduce the total number of diabetic patients in our society.

Appendix

[1] *Diabetes*. (2021, November 10). WHO | World Health Organization. Retrieved April 14, 2022, from https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] *Diabetes - Health topics*. (n.d.). WHO | World Health Organization. Retrieved April 14, 2022, from https://www.who.int/health-topics/diabetes#tab=tab_1

[3] Krishna, H. (2020, December 23). *XGBoost: What it is, and when to use it*. KDnuggets. Retrieved April 14, 2022, from https://www.kdnuggets.com/2020/12/xgboost-what-when.html

[4] *Neural Network Classification | solver*. (n.d.). Frontline Systems. Retrieved April 14, 2022, from https://www.solver.com/xlminer/help/neural-networks-classification-intro