Final Mini Project Report

**Introduction**

The data set from the CDC collected in 2014 (and accessed via UC Irvine Machine Learning Repository) contains many possible indicators of diabetes. It is important to note that in this data set, those with pre-diabetes or diabetes are both counted as having diabetes. Thus, moving forward, diabetes will refer to both those who are diabetic and pre-diabetic.

I seek to explore which controllable features have the biggest impact on whether someone is diagnosed with diabetes. Since there are some features that do help in predicting whether someone will be diagnosed with diabetes but are uncontrollable (i.e. Age, experienced a stroke, sex, etc.), I will be creating my model with all our features and then looking primarily at the controllable features (i.e. Consumption of fruits, veggies, alcohol, smoking habits, physical activity, etc.). I will follow this method because my end goal is not to predict with more accuracy who will and will not be diagnosed with diabetes, but rather to determine the most helpful practices to prevent diabetes.

**EDA**

My first exploration was to look at five number summaries of each numerical variable to make sure there were no extreme values that are unreasonable for our data set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Minimum** | **Q1** | **Median** | **Q3** | **Maximum** |
| **BMI** | 12 | 24 | 27 | 31 | 98 |
| **MentHlth** | 0 | 28 | 30 | 30 | 30 |
| **PhysHlth** | 0 | 27 | 30 | 30 | 30 |
| **Age** | 21 | 47 | 57 | 67 | 80 |
| **Education (1-6)** | 1 | 4 | 5 | 6 | 6 |
| **Income (thousands)** | 10 | 30 | 62.5 | 75 | 75 |

From these five number summaries, there do not appear to be any data entry errors as all the values are in reasonable ranges.

A graph of diabetes by age

Description automatically generatedA graph of diabetes by income

Description automatically generated

A graph of diabetes by education level

Description automatically generated

There are some interesting notes to make looking at these three graphs. We can see that the ratio of those with diabetes to those without diabetes decreases dramatically as income increases. We see a similar trend when broken down by education level, especially in the education levels 4-6. The distributions when broken up by age category look almost identical.

A diagram of a bar graph

Description automatically generated with medium confidence

This graph doesn’t necessarily give us too much insight on how physical health impacts the chance of a diabetes diagnosis, but it does tell us something interesting. This data is self-reported. We can see more data on the increments of 5. This tells us people’s reporting may not be entirely accurate to the day because people are more likely to round to the nearest interval of 5 or 10 when self-reporting. This will be important to keep in mind if we end up finding that physical health feature is one of the more important controllable features. I did not include the graph for the self-reported mental health feature, but the trend was almost identical.

**Methods**

This process began with some feature engineering. The general health variable was the only variable where the lowest category (1) was the best option. I decided to simply mirror the categories so that the order resembled the other variables. I also decided to change the category numbers for age and income to the mean of each category (i.e. Class 1 of age which was ages 18-24 is now 21, class 2 of income which was $10k - $15k is now $12.5k, etc.) The final variables I changed were MentHlth and PhysHlth. Both variables were originally a count out of the last 30 days where mental and physical health have been bad. I transformed these variables by taking their original values and subtracting them from 30. This results in the new variable as a count of the last 30 days where mental and physical health have been good. Most of these transformations were made for interpretability purposes.

**Discussion of Model Selection**

I decided to first fit a logistic regression model using all default parameters, however, the model did not converge. I upped the maximum number of iterations to 1000 which helped reach convergence while only increasing run time by about 20 seconds. The default logistic regression model parameters (other than 1000 iterations) yielded a model with about 86.4% train accuracy and 86.1% test accuracy. This was a great starting point to transition to ensemble models to allow for non-linearities in the data. I explored the following models with default parameters with the following results:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Train Accuracy** | **Test Accuracy** | **F1 Score** |
| **Random Forest** | 0.994 | 0.858 | 0.244 |
| **Gradient Boosting** | 0.867 | 0.865 | 0.249 |
| **AdaBoost** | 0.866 | 0.864 | 0.273 |
| **Extra Trees Classifier** | 0.994 | 0.852 | 0.249 |
| **KNeighbors Classifier** | ? | ? | ? |
| **Support Vectors Classifier** | ? | ? | ? |
| **Gaussian Naïve Bayes** | 0.774 | 0.772 | 0.410 |

One thing to note is most of the models have very similar test accuracies within one or two percentage points, however, the train accuracies indicate the Random Forest and Extra Trees classifiers overfit. I tried regularizing by increasing the minimum number of samples in each split from 2 to 5, 10, 20, and 30. The highest test accuracy I found was about 0.864. With this implicit regularization, we achieved about the same test accuracy.

One factor to be mentioned is the time to fit each model. The difference between most model fitting times was negligible (within a minute or two of each other), however, KNeighbors and SVC did not finish fitting after a few hours. With all these factors in mind, I decided to try a grid search to tune the hyperparameters on the gradient boosting model. I found that the best options to use were Friedman MSE for criterion, learning rate = 0.2, exponential loss, and 200 estimators. The gradient boosting model with the new parameters had an AUC of 0.828 and test accuracy of about 86.6%. With only a 0.2% increase in test accuracy, I think for the final model we will want to stick with a simpler model like logistic regression to preserve as much interpretability as possible, especially since our main goal is not prediction but to determine which factors are the most important in determining diabetes diagnosis. I also did a little bit of manual hyperparameter tuning for the logistic regression model and found the differences in test accuracy were negligible.

I examined an artificial neural network to see if there were any significant improvements in accuracy. After running a random search with 20 iterations to tune the hyperparameters, I found that 320 hyperparameters along with a learning rate of 0.001. Even after tuning the hyper parameters, the best test accuracy was 86.68%. Since there is no significant increase in accuracy, I will not use the ANN as the final model since I would lose lots of interpretability.

**Discussion of Best Model**

Based on the results above, I decided that a linear regression model would work best for our analysis purposes. I felt this was best because the purpose of my analysis is to examine which factors are most helpful to determine diabetes diagnosis. By sticking with a simpler model, we preserve the interpretation of the impact of each factor on the probability of a diabetes diagnosis.

I did, however, still want to explore the variable importance using our best ensemble method using SHAP values to cross reference with the findings later from our logistic regression model. I do not expect identical results in terms of which variables are most important compared with the logistic regression coefficients, but I do expect some crossover.

A graph of different colored lines

Description automatically generated with medium confidence

From examining these SHAP values, it appears that the top 5 factors to help determine if someone will be diagnosed with diabetes is GenHlth, Age, HighBP, BMI, and HighChol. Later in the analysis, we will see that a few of these factors are determined by the logistic regression model to be most important as well.

As mentioned above, I did some minor hyperparameter tuning with the logistic regression model. I tried L1, L2, and elastic net penalties with differing values of C which is the inverse of the strength of the regularization. I also attempted all possible solvers. There was essentially no change in the metrics with any change in hyperparameters. I decided to continue with the simple linear regression model with 1000 maximum iterations as my final model. This model yielded the following metrics:

* AUC – 0.820
* Train Accuracy – 0.864
* Test Accuracy – 0.861
* F1 Score – 0.229

Since correctly predicting diabetes was not my main goal, I knew my final model needed to have a good enough test accuracy and a decently high AUC. This means that my model has high sensitivity and false positive rate. The reason the F1 score is so low is because recall is so low. This means out of all the true positives, the model didn’t correctly identify very many. Since this was not the purpose of the analysis, we continue with the model knowing it will not predict well but will give us good insight into our factors.

**Conclusion and Next Steps**

Looking at the coefficient estimates from the model, we can start to draw some conclusions.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **HighBP** | **HighChol** | **CholCheck** | **BMI** | **Smoker** | **Stroke** | **HeartDiseaseorAttack** |
| 0.759 | 0.562 | 1.211 | 0.061 | -0.018 | 0.143 | 0.222 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PhysActivity** | **Fruits** | **Veggies** | **HvyAlcoholConsump** | **AnyHealthcare** | **NoDocbcCost** |
| -0.053 | -0.047 | -0.050 | -0.795 | 0.077 | 0.022 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **GenHlth** | **MentHlth** | **PhysHlth** | **DiffWalk** | **Sex** | **Age** | **Education** | **Income** |
| -0.540 | 0.004 | 0.008 | 0.132 | 0.259 | 0.025 | -0.031 | -0.005 |

If our question was to simply find the most important factors for diabetes diagnosis, we would determine that having a cholesterol check in the last 5 years has the biggest impact. It doesn’t make sense that having a cholesterol check in the last 5 years has any direct impact on diabetes. The causation most likely goes the other way where those with diabetes are more likely to have a cholesterol check than those without diabetes.

From there, we look at high blood pressure and cholesterol which both have coefficients far from zero. While we cannot directly change if we have high blood pressure and cholesterol or not, we can change our habits that will, in turn, raise or lower our blood pressure and cholesterol.

The other factor that has a coefficient far from zero is heavy alcohol consumption (see documentation in GitHub for what constitutes “heavy” consumption). The puzzling thing is the coefficient is negative which means the odds of a diabetes diagnosis decrease for heavy alcohol consumers compared to non/light alcohol consumers. I don’t have much domain knowledge about diabetes, but I wonder if there is something about alcohol consumption that decreases the odds of a diabetes diagnosis. Even though our model indicates that heavily consuming alcohol decreases odds of diabetes, knowing the other side effects of alcohol consumption would not cause me to advocate for heavy alcohol consumption solely to decrease chances of a diabetes diagnosis.

The final coefficient with significant weight is general health. With a coefficient of -0.540, we conclude that as your general health goes up (granted this was self-reported), your odds of receiving a diabetes diagnosis decrease.

After reflecting on these results, it does not seem that there is any one behavior measured in this study (i.e. fruit or veggie consumption, smoking, etc.) that will significantly change the odds of receiving a diabetes diagnosis. The three factors that were practically significant (as well as controllable) were high blood pressure, high cholesterol, and general health. These indicate that the chance of having diabetes is a conglomeration of our overall health rather than any one aspect. Therefore, the best way to decrease chances of diabetes is to simply live an overall healthy lifestyle.

One thing I would explore more after this analysis is the effect of alcohol on diabetes because that result surprised me.

See page 108 for age category breakdown, 23 for income breakdown, 22 for education breakdown